July 2019

Methods for Making Policy-Relevant Forecasts of Infectious Disease Incidence

Stephen A. Lauer
University of Massachusetts Amherst

METHODS FOR MAKING POLICY-RELEVANT FORECASTS OF INFECTIOUS DISEASE INCIDENCE

A Dissertation Presented

by

STEPHEN ALEXANDER LAUER

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2019

Biostatistics and Epidemiology

METHODS FOR MAKING POLICY-RELEVANT FORECASTS OF INFECTIOUS DISEASE INCIDENCE

A Dissertation Presented

by

STEPHEN ALEXANDER LAUER

Approved as to style and content by:

Nicholas G. Reich, Chair

Justin Lessler, Member

Leontine Alkema, Member

Kenneth P. Kleinman, Member

Laura B. Balzer, Member

Lisa Chasan-Taber, Department Chair
Biostatistics and Epidemiology

ABSTRACT

METHODS FOR MAKING POLICY-RELEVANT FORECASTS OF INFECTIOUS DISEASE INCIDENCE

MAY 2019

STEPHEN ALEXANDER LAUER

B.S., UNIVERSITY OF MARYLAND, COLLEGE PARK

M.S., UNIVERSITY OF MASSACHUSETTS, AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Nicholas G. Reich

Infectious diseases place an enormous burden on the people of the developing world and their governments. When, where, and how to allocate resources in order to slow the spread of a virus or deal with the aftermath of an outbreak is often the responsibility of local public health officials. In this thesis, we develop statistical methods for forecasting future incidence of infectious diseases and estimating the effects of interventions designed to reduce future incidence, bearing in mind the needs and concerns of those public health officials.

While most infectious disease forecasting models focus on short-term horizons (*i.e.* weeks or months), long-term forecasts made prior to the epidemic season may be more useful to public health officials. In Chapter 2, we make an annual forecasting model for dengue hemorrhagic fever incidence based on early season incidence, weather, and

demographics. The predictions from this forecasting model outperform a baseline model based on the ten-year median on out-of-sample data. To our knowledge, this model makes accurate annual forecasts earlier in the year than any other dengue model on record.

After public health officials implement an intervention, whether a preventative action or a response to a developing outbreak, they may want to know whether that intervention was effective. In Chapter 3, we evaluate an effect estimation technique, called covariate-adjusted residuals, within a causal inference framework. This technique was originally developed for use in randomized trials, but has also been used in observational settings in ecology. Much research in the field of causal inference has focused on developing methods that account for confounding in non-randomized experiments. To our knowledge, we are the first to evaluate covariate-adjusted residuals from a causal inference perspective, and to develop an extension for use in observational studies.

In Chapter 4, we investigate whether using forecasts can improve the efficacy of effect estimation. In certain situations, forecasting can be used for covariate selection and dimension reduction that improves the performance of covariate-adjusted residuals in estimating the effect of an intervention. We used our findings to estimate whether an intervention for Zika reduced dengue hemorrhagic fever incidence in Thailand in 2016.

CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INFECTIOUS DISEASE FORECASTING FOR PUBLIC HEALTH

(The contents of this chapter are under revision for the book *Population Biology of Vector-borne Diseases*, co-authored with Alexandria C. Brown and Nicholas G. Reich.)

"... diviners employ art, who, having learned the known by observation,
seek the unknown by deduction."
– Cicero (as quoted in [123])

"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."
– LaPlace (1825) (as quoted in [182])

1.1 Background

1.1.1 A brief history of forecasting

The ability to foretell, or divine, future events has for millennia been seen as a valued skill. While there are records of Babylonians attempting to predict weather patterns as early as 4000 BCE based on climatological observations [130], early attempts at

divination were just as likely to be driven by unscientific observation. However, in the last 150 years, rapid technological advancements have made data-driven forecasting a reality across a number of scientific and mathematical fields.

The science of forecasting was pushed forward especially in the second half of the 20th century by the fields of meteorology and economics, but more recently other fields have started to build on this research. Examples include world population projections [64, 155], political elections,[29, 69, 112], seismology[54, 25], as well as infectious disease epidemiology [15, 115, 199, 81, 158].

Forecasting has been an active and growing area of research for over a century (1.1), with particular acceleration observed since 1980. While research focused on forecasting infectious diseases started in earnest in the 1990s, since 2005 the number of articles on infectious disease forecasting has increased seven-fold, at a faster pace than research on general forecasting during that time, which increased by a factor of 3. In 1991, forecasting was the topic of one of every thousand published academic papers, based on counts from the Science Citation Index and the Social Science Citation Index, obtained from the Web of Science. In 2017, despite an overall rise in academic publication over previous decades, over four of every thousand published papers were about forecasting.

1.1.2 What is a forecast?

In common parlance, there is not a strong distinction between the terms 'prediction' and 'forecast.' Nor does there exist a strong consensus in the biomedical, ecological, or public health literature on the distinction. Nate Silver has suggested that etymologically, the term forecast "implied planning under conditions of uncertainty" in contrast to prediction, which was a more ancient idea associated with superstition and divination [182]. In the modern scientific world, some fields, such as seismology, use the term forecast to refer to a probabilistic statement in contrast to a prediction

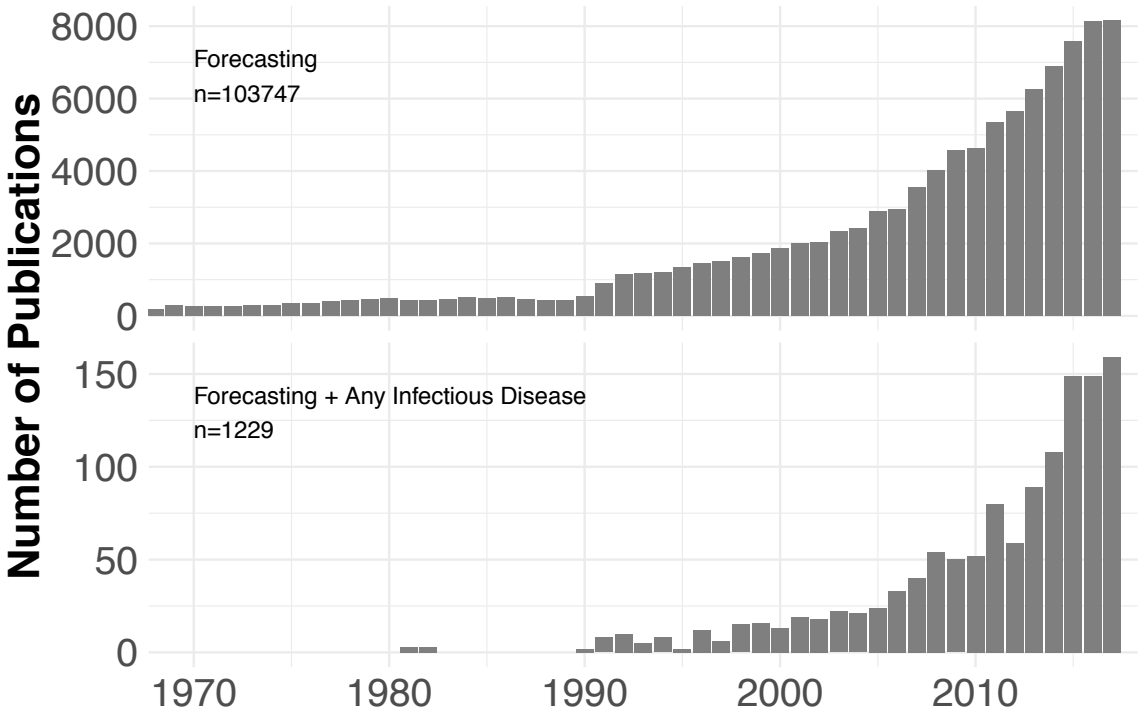

Figure 1.1. Publication trends from 1970 through 2016. The y-axis in each panel shows the number of publications, according to the Web of Science, for papers with the topic of (A) 'forecast*', and (B) 'forecast*' + any of a list of infectious diseases taken from WHO [206]. There are 1,989 and 0 publications, respectively, that were published prior to 1970. All counts are taken from the Science Citation Index and the Social Science Citation Index, obtained via the Web of Science database.

which is a "definitive and specific" statement about a future event. In other fields, the difference in meaning is even less clearly defined, with forecasting often connoting the prediction of a future value or feature of a time-series [40]. We note that convention in biomedical research uses the term 'prediction' to refer to an individual-level clinical outcome (*e.g.* risk-scores give individualized predictions of heart attack or stroke risk), but both 'prediction' and 'forecast' are used interchangeably to refer to events or outcomes that may impact more than one person. The term 'forecast' may often refer more broadly to a trend or event observable or experienced by many individuals rather than a local, individualized outcome. Our use of the term forecasting will take aspects from several of these definitions. Specifically, we define a forecast as *a quantitative, probabilistic statement about the uncertainty surrounding an event, outcome, or trend that has not yet been observed, conditional on data that has been observed.*

Note that a forecasted event or outcome need not necessarily be in the future, as events in the past that have not yet been quantitatively measured may also be forecasted. For example, on May 1 we may have data for a particular time-series available through April 1. We could make a "forecast" of the time-series for April 15 even though this event is in the past. This type of forecast has been referred to as a "nowcast", although more generally, this is a special case of a forecast.

1.1.3 Forecasting challenges that are specific to infectious disease

There are operational and statistical challenges in forecasting that are specific to the setting of infectious disease. These challenges in and of themselves may not be unique to the field, but taken together, they describe obstacles that forecasters face when taking on a problem in infectious disease.

1.1.3.1 Challenge 1: Unobserved Complexity

When attempting to forecast the transmission of an infectious disease, researchers need to account for processes on scales from micro to macro. Behaviors of and

interactions between viruses, vectors, hosts, and the environment each play a part in determining the spread of a disease. Fundamentally, viruses are microscopic organisms whose interactions with their hosts (humans or vectors), competition with other viruses, and genetic mutations determine their potential for survival, transmission, and the severity of their induced infections. Whether biological data can be used for forecasting population-level transmission remains to be seen (Section 1.2.3.1). Researchers have developed mechanistic models based on biological and behavioral principles that encode the processes by which diseases spread between humans (Section 1.2.1.1). Environmental conditions may dictate not only the life cycle of vectors, such as *Aedes* mosquitoes, but also human behavior. The inclusion of climate covariates has a mixed record for forecasting disease transmission (Section 1.2.3.2).

Therefore the key challenges for forecasting infectious disease are both in the complexity of the biological and social models needed, but also in the available data [133]. For forecasting weather, scientists rely on thousands of sensors across the world, collecting continuous real-time data. These rich and highly accurate data streams are unavailable to infectious disease researchers whose gold-standard data by and large come from very human systems: epidemiological surveillance systems that capture only a fraction of all cases and often are reported with substantial delays.

1.1.3.2 Challenge 2: The Forecasting Feedback Loop

Forecasts of disease incidence can encourage governments and public health organizations to intervene in order to slow transmission. If forecasts of infectious disease are used to inform targeted interventions or risk communication strategies and the interventions change the course of the epidemic, then the forecast itself becomes enmeshed in the causal pathway of an outbreak. This feedback loop has been identified as the single most important challenge separating infectious disease forecasting from forecasting natural phenomena such as weather [133].

In settings where forecasts will be used to inform interventions, this feedback loop of infectious disease forecasting should be taken into account in the forecasts. Without such accounting, if a forecast predicts an outbreak and triggers an intervention that prevents the epidemic from taking off, then the forecast itself would be seen as wrong, despite this being a public health victory. This implies that forecasting models should, when in these settings, create multiple forecasts under different intervention scenarios. Mechanistic forecasting models, that use explicit disease transmission parameters, may be best suited for these types of forecasts, since intervention effects could be incorporated directly as impacting these parameters.

Method development is needed in this area to address open scientific questions. What model frameworks can best balance forecast accuracy with the ability to incorporate multiple potential future scenarios? Can forecast models be used to assess intervention effectiveness?

1.1.4 Definitions and basic notation

Here, we will introduce some basic mathematical notation for time-series forecasting that we will use throughout this chapter. In many forecasting applications, including many of those for infectious disease, the available data are often a time series of observed values for a particular location or setting. For infectious disease applications, these data are often a measure of incidence, such as case counts or the percentage of all doctor visits with primary complaint about a particular disease. In the text that follows, we will use language specific to that of spatio-temporal disease incidence data, although much of what we describe can be applied more generally as well.

1.1.4.1 Data

We will start with a toy example and later extend the notation to more realistic scenarios. In our example, we have a complete (*i.e.* no missing data) time series of infectious disease case counts from a single location, such as a school or hospital. We

define y_t as an observed value of this incidence in time interval t from our time series $\{y_1, y_2, y_3, \ldots\ y_t, \ldots, y_T\}$. We assume that these observations are draws from random variables $Y_1, Y_2, Y_3, \ldots, Y_t, \ldots, Y_T$, whose probability distributions can be thought of as a function of t, prior values of y represented as $y_{1:t}$, and other covariates x_t. We use T throughout to refer to the total number of time points in the time-series and t to refer to a specific time point relative to which a forecast is generated.

Two important features of our observed data are frequency and scale. In our example, incidence is recorded at regular time intervals. Furthermore, many processes, including our infectious disease time series, have a cyclical element. We define the frequency of a time series as the number of observations within a single cycle. For example, if we have monthly incidence data and know that there are annual weather patterns that influence incidence in our observed data, the frequency of our time series would be 12 months/observations.

1.1.4.2 Targets

Targets are the as-yet-unknown features of the data that are the subject of forecasting. In our toy example, we may want to forecast incidence at a certain future time—but targets can be a variety of endpoints extrapolated from the observed data. For forecasts of the time-series values itself, *i.e.* when a target is defined to be a past or future value of the time-series Y_{t+k}, we use a special nomenclature, referring to them as 'k-step-ahead' forecasts. We define $Z_{i|t}$ as a random variable for target i positioned relative to time t. For example, in the infectious disease context, $Z_{i|t}$ could be:

- incidence at time t, or Y_t,
- incidence at time $t + k$ either in the future or past relative to time t, or Y_{t+k}, where k is a positive or negative integer,

- peak incidence within some period of time or season, or $\max_t(Y_t)$ where t are defined to be within a season,
- the time at which a peak occurs within some season, or $\{t' : Y_{t'} = \max_t(Y_t)\}$
- a binary indicator of whether incidence at time $t+k$ is above a specified threshold C, or $\mathbb{1}\{Y_{t+k} > C\}$.

1.1.4.3 Forecasts

A forecast, as defined in Section 1.1.2, must provide *quantitative and probabilistic* information about an outcome. In the context of this notation, a forecast can be represented as a predictive density function for a target, or $f_{z_{i|t}}(z|y_{1:t}, t, x_t)$. The form of this density function will depend on the type of variable that Z is, and it could be derived from a known parametric distribution or specified directly. For example, if the target is a binary outcome (*e.g.* whether in week 4, the observed incidence will be above 10 cases) the density could be specified as a Bernoulli distribution with a parameter associated with the probability of the outcome occurring. It could also be specified directly as a probability that the incidence is >10 and the probability that the incidence is ≤ 10. For a continuous target (*e.g.* the number of new cases occurring in February), the predictive density could be represented by a Poisson distribution with a given mean or as a vector of probabilities associated with all possible integer values of cases.

To enable clear definitions for forecasting in real-time, forecasts must be associated with a specific time t. This time t represents the point relative to which targets are defined. For example, if a forecast is associated with week 45 in 2013, then a '-1-step-ahead' (read 'minus-one-step-ahead') forecast would be associated with incidence in week 44 of 2013 and a '3-step-ahead' forecast would be associated with week 48.

1.1.4.4 Forecast time-scale

Another consideration for infectious disease forecasting is the forecast horizon, the temporal range that the forecast predicts [137, 184]. Regardless of the model type, many recent infectious disease forecasting efforts have focused on short time scales (weeks or months) [17, 27, 63, 83, 114, 120, 142, 158, 176, 178, 189, 211]. These studies demonstrated the importance of recent case counts and seasonality on the immediate trajectory of infectious disease incidence. In 2015, the National Oceanic and Atmospheric Administration (NOAA) and the Centers for Disease Control and Prevention (CDC) hosted a competition to make within-season forecasts for longer forecast horizons, such as annual dengue incidence, epidemic peak, and peak height, for San Juan, Puerto Rico and Iquitos, Peru [138]. Prior to these competitions, long-term forecasts were more commonly used for chronic disease prevalence than for non-chronic infectious disease incidence [184].

1.2 Models used for forecasting infectious diseases

1.2.1 Mechanistic vs. Statistical: a taxonomy of forecasting models

According to Myers *et al.*, forecasting models for infectious diseases take either a 'biological approach' or a 'statistical approach' [137]. Others have phrased this distinction as one of mechanistic (*i.e.* biological) and phenomenological (*i.e.* statistical) models. A model based on disease biology can account for previously unforeseen scenarios that are possible due to transmission dynamics, however these models often require specification of a large number of parameters and covariates in order to make forecasts. On the other hand, statistical forecasting models are restricted by the assumption that future incidence will follow the patterns of incidence observed in the past, but can be specified without full knowledge of the disease process or interactions between members of the population. In this section, we will discuss the major modeling methods across the biological-statistical spectrum (1.1).

Table 1.1. Taxonomy of models and methodologies for infectious disease forecasting.

Methods for mechanistic models	**Methods for phenomenological models**
Deterministic differential equations	Generalized linear models
Stochastic differential equations	- climate and/or AR terms
Chain binomial models	- sinusoidal seasonality
Agent-based simulation models	- penalized regression (*i.e.* "large p")
Filtering approaches (Kalman, particle)	ARIMA and extensions
Bayesian networks	Classification and Regression Trees
TSIR methods	Kernel Conditional Density Estimation
Growth models	Ensemble methods

1.2.1.1 Mechanistic models

Compartmental models are the standard biological, or mechanistic, approach for modeling infectious disease [102, 181, 111]. Kermack and McKendrick proposed the first such model, now known as the susceptible-infectious-recovered (SIR) model, in which members of a population transition through each compartment (susceptible to infectious to recovered) over the course of an epidemic [105]. While this process mimics the behavior of an outbreak, the simplest model assumes that the population is "well mixed", such that each individual is equally likely to encounter any other individual. Since this is unlikely, researchers can add more compartments (*e.g.* adults and children), along with contact rates between compartments, or individually model each member of the population (*i.e.* agent-based modeling) [49]. Though greater complexity requires more modeling assumptions, compartmental models have been effective at estimating underlying infectious disease processes and the potential impact of interventions [111, 52, 102, 160]. Recently, several papers extended compartmental models for use in forecasting infectious disease incidence, particularly for influenza [17, 142, 145]. Additionally, Shaman and Karspeck incorporated humidity into a SIRS compartmental model to forecast influenza [175, 176]. And SIR models have been used to forecast dengue outbreaks [209].

1.2.1.2 Classical statistical models

On the statistical side of the modeling spectrum, many regression-style methods have been used for forecasting. Perhaps the most well-known statistical method for time series is the auto-regressive integrated moving average, or ARIMA [22]. ARIMA models use a linear, regression-type equation in which the predictors are lags of the dependent variable and/or lags of the forecast errors. ARIMA and seasonal ARIMA (SARIMA) models are frequently applied to infectious disease time series [94, 181, 184, 193, 156]. Lu *et al.* combined a SARIMA model with a Markov switching model (a type of compartmental model) to account for anomalies in the surveillance process [120].

Also under the subheading of trend and seasonal estimation are simple exponential smoothing strategies, known as Holt-Winters models [86, 204]. Exponential smoothing techniques involve taking weighted averages of past observations with exponentially decreasing weights further from the present. Holt-Winters in particular is known for its efficient and accurate predictive ability [62, 68]. These approaches have been used successfully in forecasting dengue fever [26] and leprosy [37].

Some researchers have used generalized linear regression models to develop infectious disease forecasts. In some cases, researchers used lagged covariates (*e.g.* temperature, rainfall, or prior incidence) to predict future incidence [75, 83, 114, 132, 158]. Held and Paul also combined statistical and biological theory by building a regression model that consisted of three components of disease incidence: endemic, epidemic, and spatio-temporal epidemic (to account for spread of disease across locations) [80]. This has become a well-established framework for forecasting infectious disease surveillance data [85, 81, 156], and is accompanied by software implementing the methods [129].

1.2.1.3 Modern statistical methods

Modern statistical methods, *i.e.* not the classical time-series and regression-based approaches, are an increasingly popular way to forecast infectious disease incidence. These methods include non-parametric approaches as well as more black-box machine-learning style algorithms. We focus in this section on stand-alone forecasting methods, for a discussion on ensemble methods, see Section 1.2.4.

Several papers have found that machine-learning modeling methods can outperform standard statistical models for infectious disease forecasting: random forests outperformed ARIMA forecasting avian influenza [100], a maximum entropy model outperformed logistic regression forecasting hemorrhagic fever with renal syndrome [113], and fuzzy association rule mining outperformed logistic regression forecasting dengue outbreaks [27]. Additionally, kernel conditional density estimation, a semi-parametric method, was shown to have more well-calibrated probabilistic forecasts than SARIMA and other regression-based approaches for forecasting highly variable dengue outbreaks in San Juan, Puerto Rico [156]. Neural networks have also been used for forecasting influenza [208] and Hepatitis A [72].

1.2.1.4 Comparisons between mechanistic and statistical models

From an epidemiological perspective, mechanistic models have several clear advantages over statistical models. They are biologically motivated, and therefore have parameters that relate to well-established theory and can be interpreted by experts in the field. Mechanistic models can flexibly incorporate features such as interventions or behavioral changes, which can be critical, especially if forecasts are desired for different intervention scenarios (see Section 1.1.3). While mechanistic models can be built to rely heavily on previously observed data, they also can be instantiated with very little prior data, such as in emerging outbreaks (see Section 1.2.2). Additionally, while forecasts from statistical models are typically bounded by trends that have been

previously observed, mechanistic models can comfortably forecast outside of previously observed trends if the underlying states of the model call for such dynamics.

Despite these advantages, in forecasting settings where substantial historical data is available, statistical models may prove more effective at using past observed trends to forecast the future. Many statistical models were designed to be either more flexibly or parsimoniously parameterized, meaning that they may be able to more easily capture dynamics common to infectious disease time-series such as auto-regressive correlation and seasonality. Additionally, they can be built to rely less heavily on specific assumptions about a particular biological disease transmission model, giving them flexibility to adapt when the data does not follow a particular pattern. In other words, since any specified mechanistic model is necessarily a simplification of the true underlying disease process, the question is how much will forecast accuracy suffer as a result of the inevitable model misspecification. In many cases, heavily parameterized mechanistic models may be more sensitive to model misspecification than a more flexible statistical model.

Despite many unanswered questions about when and in what settings one type of model will generally do better than the other, research that make explicit, data-driven comparisons are fairly uncommon. Multi-team infectious disease forecasting challenges provide some of the best data available on this important question. For forecasting seasonal influenza, what limited data there are suggest that mechanistic and statistical approaches show fairly similar performance, with statistical models showing a slight advantage [125, 157, 93]. A collaborative effort during the 2014 West Africa Ebola outbreak to forecast synthetic data showed fairly comparable results from mechanistic and statistical models and did not make an explicit comparison between the two [198]. Summary analyses from other challenges have not been published, but we note that a very simple quasi-mechanistic model was the best performing model in forecasting the pattern of emergence of Chikunguya in the Americas [110, 36]. In sum, more research

is needed to improve our understanding about whether different types of forecasting methods can be shown to be more reliable than others in certain situations.

1.2.2 Forecasting in emergent settings

In emerging outbreak scenarios, where limited data is available, mechanistic models may be able to take advantage of assumptions about the underlying transmission process, enabling rudimentary forecasts even with minimal data. On the other hand, many statistical models without assuming a mechanistic structure rely on past data to be able to make forecasts. That said, any forecasts made in settings with limited data must be subjected to rigorous sensitivity analyses, as such forecasts will necessarily be heavily reliant on model assumptions.

A wide range of different mechanistic models have been used in settings where infectious disease forecasts are desired for an emerging threat. A simple non-linear growth model performed the best in a prospective challenge for forecasting Chikungunya in the Americas [110]. Unlike many other mechanistic forecasting approaches, this model has a small number of parameters, is easy to fit, and makes only a few assumptions about the underlying disease process. A deterministic SEIR model was used to forecast synthetic Ebola epidemic data, showing comparable performance to other methods on the same data [59, 199]. A stochastic SEIR model also forecasted synthetic Ebola data, and showed somewhat less reliable performance compared to other methods [57]. Data-driven agent-based models have also been shown to be a viable forecasting tool for emerging infectious diseases [197]. An SIR model, similar to one used to forecast seasonal dengue fever and influenza, was used with a more complex compartmental structure to forecast the spread of Ebola during the outbreak in West Africa in 2014 [178]. A set of quasi-mechanistic models were used to forecast Zika virus transmission during 2017 to help plan for vaccine trials [7].

1.2.3 Using external data sources to inform forecasts

1.2.3.1 Moving beyond surveillance data

Traditional approaches to infectious disease forecasting often have relied on a single time-series, or multiple similar time-series (*e.g.* incidence from multiple locations). However, other types of epidemiological data may provide important information about current transmission patterns.

Leveraging laboratory data, collected either through passive or active surveillance strategies, may provide crucial data about what specific pathogens are currently being transmitted and could inform forecasting efforts. This is an area that warrants more research, as few efforts have tackled the challenge of having laboratory test data inform forecasts at the population level. The most progress has been made in forecasting influenza transmission at this level. One model uses an aggregate measure of genetic distance of circulating influenza strains from the strains in the vaccine as a variable to help forecast peak timing and intensity of seasonal outbreaks in the US [43, 44]. Some efforts have also been made to make strain-specific forecasts for influenza [99]. Other efforts have focused on longer-term forecasts of what strains will predominate in a given season, with an eye towards providing information to influenza vaccine manufacturers [134]. These efforts have moved beyond influenza, and forecasting pathogen evolution is being worked on for a variety of different pathogens [74].

Another, and very different, kind of epidemiological data for forecasting is expert opinion. Long seen as a useful indicator in business applications [190], expert opinion has recently begun to be used in infectious disease applications [51, 37]. While not traditional clinical data, expert opinion surveys leverage powerful computers, *i.e.* human brains, that can synthesize historical experience with real-time data. Intuitive interfaces can facilitate the specification of quantitative and digitally entered forecasts from experts who need not be technically savvy, lowering the barriers to participation and subsequent analysis [51]. In the 2016/2017 influenza season in the

US, a forecast model based on expert opinion was a top-performer in a CDC-led forecasting competition (CDC FluSight Presentations, 2017). Human judgment and expert opinion surveys are a promising area for further forecasting research, especially in contexts with limited data availability.

1.2.3.2 Digital epidemiology

Digital epidemiology has been defined as the use of digital data for epidemiology when the data were "not generated with the primary purpose of doing epidemiology" [170]. Broadly speaking, this might refer to online search query data, social media data, satellite imagery, or climate data, to name a few. These resources may hold promise for forecasters who want to incorporate "Big Data" streams into their models. In the past 10 years, much research has explored the potential for leveraging multiple data streams to improve forecasting efforts, but this practice is still in its nascent stages. So far, the utility of digital epidemiological data for forecasting has been somewhat limited, perhaps due to challenges in our understanding of how digital data generated by human behavior and interactions with the digital world relate to epidemiological targets [133, 152, 170].

Perhaps the most famous and controversial example of using digital data streams to support infectious disease prediction surround the early promising performance[65, 45] and later dismal failure [109] of Google Flu trends to predict the influenza-like-illness in the US. Google Flu trends was based on tracking influenza-related search terms entered into the search engine. Although Google eventually discontinued the public face of the project due to poor performance, criticism of the Google Flu trends approach centered around how data was included or excluded, interpreted, and handled rather than the algorithm that produced the actual forecasts [172, 144]. Ongoing research on using search engine data in forecasting has continued despite the failure of Google Flu trends, producing incremental but consistent improvements to forecast accuracy

[212, 213, 124, 119]. To date, there has been less research investigating whether using real-time search data could improve the timeliness of forecasts and nowcasts in settings where reporting delays limit the utility of real-time surveillance data (see Section 1.4.1).

The use of climate data for epidemic forecasting serves as another clear example of repurposing data for epidemiology. While climate factors are known biological drivers of infection risk (*e.g.* the impact of absolute humidity on influenza virus fitness [177], or temperature and humidity providing optimal conditions for mosquito breeding), the evidence supporting the use of climate data in forecasting models is mixed. Climatological factors such as temperature, rainfall, and relative humidity were used to forecast annual counts of dengue hemorrhagic fever in provinces of Thailand (to be seen in Chapter 2).However, only temperature and rainfall were included after a rigorous covariate selection process and neither were included in the final model, although subanalyses showed variation in these associations across different geographic regions of Thailand. Climate factors were shown to improve forecasts of dengue outbreak timing in Brazil[118], but played a less influential role in dengue forecasts in Mexico [94]. Aggregated measures of absolute humidity have been incorporated into influenza forecasts in the US [176, 212]. However, without clear standardization across these studies, these mixed results may reflect heterogeneity in the spatial scales at which forecasts are made, and in the spatial and temporal scales at which climate factors are measured, are aggregated, and drive disease transmission.

1.2.4 Forecasting with ensembles

Ensemble forecasting models, or models that combine multiple forecasts into a single forecast, have been the industry standard in weather forecasting for decades. By fusing together different modeling frameworks, ensembles that have a diverse library of models to choose from end up incorporating information from multiple

different perspectives and sources. When appropriate methods are used to combine the forecasts, the resulting ensemble should in theory always have better long-run performance than any single model.

Ensembles have been increasingly used in infectious disease applications and have shown promising results. For forecasting influenza, several model averaging approaches have shown improved performance over individual models [210, 156]. Similar approaches have yielded similar results for dengue fever [209] and Ebola [199].

In many of these examples, however, the number and diversity of distinct modeling approaches was fairly small. To unlock the full potential value of ensemble forecasting, as well as understanding the added value of contributions from new and different data sources or modeling strategies, more scalable frameworks for building forecast models are required. There is a need to develop infrastructure and frameworks that can facilitate the building of ensemble forecast models. This will require clear technical definitions of modeling and forecasting standards.

1.3 Components of a Forecasting System

Due to the complex biological, social, and environmental mechanisms underlying infectious disease transmission, the data generating process that we are attempting to model is often unobservable. That means that we have many potential models to choose from. How do we decide which model is the best for forecasting future targets? We need a forecasting system that specifies how we plan to build our models, make our forecasts, and evaluate the results in order to determine which model performs best in our given situation. Because models perform differently depending on the forecast target, type of forecast, model training technique, and evaluation metric, it is important to specify the forecasting system prior to fitting the models to ensure an optimal model selection [6].

1.3.1 Forecast type

When building a forecasting system, the first steps are to choose the forecast target (as described in Section 1.1.4) and type. Forecast targets are often dictated by the goals of a public health initiative. Researchers and public health officials collaborate to find a forecast target that is most useful for allocating resources and implementing interventions to reduce the severity of an infectious disease outbreak. The forecast target helps inform the selection of the forecast type, of which there are three: point, interval, and density forecasts [40].

A point forecast is a forecast of a single value that attempts to minimize the error between that value and the eventually observed value. The mean, median, or mode of a predictive distribution is often used as the point forecast for a specified target. While point forecasts are simpler to produce and interpret, they may make inaccurate assumptions about the underlying probability distribution, leading to low-quality forecasts. For example, a point forecast based on the mean may represent a value for which there is actually a small likelihood of occurring (between the peaks of a multi-modal distribution). This could mislead officials and researchers into forecasting a medium-sized outbreak when the full distribution actually shows that the most likely future scenarios are for either low incidence or an epidemic outbreak.

Interval forecasts supplement point forecasts with a range of likely values. The level of a prediction interval indicates the percentage of eventually-observed outcomes that should fall within that interval; *i.e.* if a model makes 100 forecasts, about 95 should fall within the 95% prediction interval. Interval forecasts can be produced as simply as adding symmetric bounds on either side of the point interval (often determined parametrically) or by using more complex methods such as non-parametric bootstraps and Bayesian posterior distributions.

Density forecasts assign probabilities to all possibly observed values to form a distribution from which an interval or point forecast could be derived. The goal

of a density forecast is to assign the maximum probability to the true future value. Density estimation often requires simulation-generating methodology, which can be more time-consuming and computationally-intensive than other techniques. Ongoing advances in computing continue to make density forecasting methods more feasible for researchers. Density forecasts contain the most nuanced information of all of the forecasting methods, but are often the most difficult to interpret and communicate to non-expert collaborators. This type of forecast may require further analysis or interpretation to provide meaningful information to public health officials.

1.3.2 Evaluation and scoring

Armstrong [6] and Hyndman and Koehler [88] list a number of desirable features for scoring metrics, especially for point forecasts. Each metric has both strengths and weaknesses in different forecasting contexts. Research suggests that metrics should be *scale-independent and insensitive to degree of forecast difficulty*. Oftentimes, observations within an infectious disease time series have a Poisson distribution, in that variability increases with the expected value of an observed value. Thus, incidence near the seasonal peak are both larger and more variable than incidence near the seasonal nadir and, consequently, forecasting model error will depend on the size of the value it is forecasting. In these situations, using logged metrics can weight errors equally across different scales [159].

Furthermore, *metrics should be defined and finite in reasonable scenarios*. Some metrics may be undefined or infinite due to division by zero and sometimes this can reveal issues with a model. However, if this happens in inappropriate contexts the metric loses its utility. A metric should be *valid* in that it should agree with both experts in the field (face validity) and most other metrics (construct validity). For instance, even non-experts can agree that a model that forecasts negative values of disease incidence should be considered invalid. Since forecasting model performance

varies across metrics, we would prefer to use one that is in general agreement with other metrics as opposed to an outlier. When forecasting, we should use an *unbiased* metric that should not reward forecasts that are above (below) the target more than those that are below (above) the target. Asymmetric cost functions can help officials decide a course of action, however that is a separate task from forecasting disease incidence.

For point forecasts, the current best practice is to use a metric that scales the forecasting error against that of a reference model [88, 159, 66]. One example is the relative mean absolute error, which divides the mean absolute error of the forecasting model by the mean absolute error of a reference model $\left(\text{rMAE} = \frac{1/n\sum_{t=1}^{n}|y_t - \widehat{Y}_t^{\text{forecast}}|}{1/n\sum_{t=1}^{n}|y_t - \widehat{Y}_t^{\text{reference}}|}\right)$. An additional desirable feature that this metric has is interpretability, in that $rMAE < 1$ means that the forecasting model has less error than the reference model and $rMAE > 1$ means that the forecasting model has more error than the reference model.

Interval forecasts can be evaluated by their coverage rate and their width; prediction intervals should be as narrow as possible while covering a proportion of forecasts approximately equal to that expected by its level. Gneiting and Raftery [66] describe a useful interval evaluation metric:

$$S_{\alpha}^{\text{int}}(y_t, u_t, l_t) = \frac{1}{T}\sum_{t=1}^{T}(u_t - l_t) + \frac{2}{\alpha}(l_t - y_t)\mathbb{I}(y_t < l_t) + \frac{2}{\alpha}(y_t - u_t)\mathbb{I}(y_t > u_t)$$

where l_t and u_t are the lower and upper bounds of a $(1-\alpha) * 100\%$ prediction interval for observation y_t and we would like to minimize the score S_{α}^{int}. Forecasting models are penalized for having wider intervals and for having observed values that fall far outside of the intervals. Observations that fall outside of large prediction intervals (small α) are penalized more than those that fall outside of small prediction intervals (large α).

Gneiting and Raftery state that a probabilistic forecast should have a distribution that is consistent with the distribution of the observed values and that models that assign more weight to the eventually observed values are better than those that do not [66]. Metrics that provide incentives to meet these two criteria are "proper scoring rules". Gneiting and Raftery propose a number of proper scoring rules for many different situations. A commonly used proper scoring rule is the log score (LogS $= 1/n \sum_{t=1}^{n} \log P(\widehat{Y}_t = y_t)$), which goes to 0 if all of the probability is correctly placed. However, this method may be sensitive to outliers, as any observation with a forecasted probability of zero causes the metric to go to negative infinity (though adjustments can be made to avoid this). As an alternative, Funk *et al.* recommend using multiple metrics to evaluate the unbiasedness, calibration, and sharpness of infectious disease forecasts [58].

The continuous ranked probability score (CRPS) is a proper scoring rule that measures the difference between the forecasted and observed cumulative distributions [82]. This metric measures both the bias and the uncertainty of the forecasted density and thus rewards forecasts that assign weight closer to the observed value, even if it doesn't assign much weight exactly on the observed value. A point forecast with no uncertainty will have a CRPS equal to the absolute error of the forecast. Unbiased forecasts with more uncertainty will have a higher CRPS than for unbiased forecasts with less uncertainty, however biased forecasts with more uncertainty can have a smaller CRPS than biased forecasts with less uncertainty. While CRPS is scale-dependent, dividing the CRPS of a forecasting model by the MAE of a benchmark model (as in the relative mean absolute error) yields a scale-independent continuous ranked probability skill score [23, 20].

1.3.3 Model training and testing

In order for a forecasting model to be useful for researchers or officials it needs to be generalizable to data beyond the observations that were used for fitting. For instance, a dengue model that perfectly forecasts monthly observations over the past ten years, but performs worse than a reasonable guess—*e.g.* the average monthly incidence—over the next five years is not very useful. We would be better off using the reasonable guess instead of the forecasting model. Though we can never be sure that our best model will perform well outside of our dataset, we can get a better idea of its *out-of-sample* performance using model training and testing. We will illustrate this concept with an example from Chapter 2,in which we forecasted annual dengue hemorrhagic fever (DHF) incidence in Thailand for 76 provinces.

Prior to fitting any model, we split our data into a 'training' sample (for initial model selection) and a 'testing' sample (for final model evaluation) [187, 78]. The training sample is used for model experimentation and parameter tuning. The testing sample is sequestered until we are ready to characterize the performance of our chosen model in the final analysis. Why do we do this? Models tend to "overfit" to the data that is used for estimating parameters, meaning that the best set of parameters for one set of data are often not generalizable to other data. For instance, when using least squares to fit a model, adding a new covariate – no matter how arbitrary – will always decrease the error of the fitted model; *i.e.* the model residuals are always smaller in a model with more components. We would prefer a model-selection method that minimizes the error on the testing sample, which would be more generalizable to new observations. In our example, we split the data so that the first 10 years (760 observations) were for training and the last 5 years (380 observations) were for testing.

Our next challenge is to train our model in such a way that we minimize the error on the testing data – without using the testing data! There are many methods of doing so, the most popular of which are using information criterion (AIC or BIC) or

by sampling the training data (cross-validation or bootstrapping). These methods each have strengths and drawbacks, the details of which are outside the scope of this work. For further reading, the authors recommend *The Elements of Statistical Learning* [78]. Bergmeir, Hyndman and Koo (2017) investigated the performance of time-series specific cross-validation methods [12]. For our purposes, it is important to note that these methods still tend to reward slightly more complex models that may have more error on the testing data than a smaller model would [179]. Thus, in addition to selecting the model that performs best by our pre-specified information criterion or cross-validation metric, we should choose a more parsimonious model that has more error in the training phase as a check against overfitting [141].

In our example, we ran leave-one-season-out cross validation on the training phase data to select our model. In this procedure, we fit a model on 9 of the 10 years to predict the final year — *e.g.* fitting on 2001-2010 to predict 2000. We repeated this to predict the provinces in each of the 10 years, recorded the error for each prediction, and then took the mean absolute error across all predictions and called it the "cross-validation (CV) error" for a given model. We performed cross-validation for 202 models with different specifications and covariate combinations. The model that minimized the CV error had 5 covariates, while the model that minimized the residual error across the entire training phase had 14 covariates (Figure 1.2). In addition to the 5-covariate model, we also selected the smallest model within one standard deviation of the smallest CV error — a univariate model — to forecast the testing phase.

We use the two models that we selected in the training phase to forecast the testing phase. How this is executed depends on the goals of model evaluation and the features of the model fitting process. There are several "windows" used for making testing phase forecasts as outlined by Bergmeir and Benítez [11], of which we'll highlight two. When using the *rolling-origin-recalibration* window, we fit the model to the training data to forecast the first testing phase observation, then we move the first

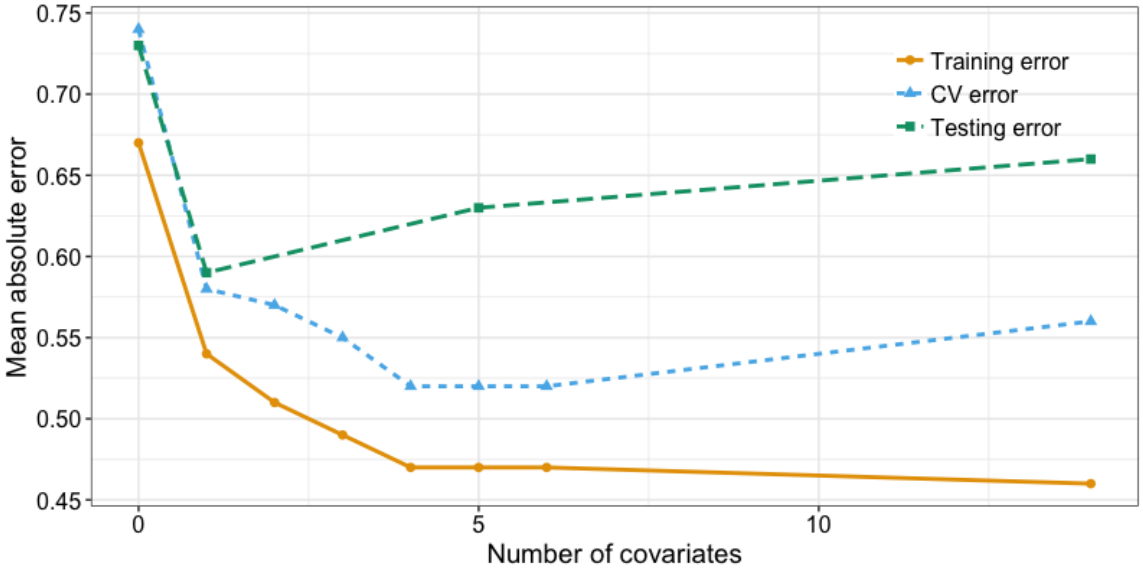

Figure 1.2. The training error (orange, solid line), cross-validation (CV) error (blue, dotted), and testing error (green, dashed) by number of model covariates for an applied example. The training error is monotonically decreasing as the number of covariates increases. The CV error is minimized at 5 covariates and better approximates the testing error than the training error, especially for fewer covariates. The univariate model had the least error in the testing phase.

observation from the testing phase into the training data, re-fit the model and forecast the second testing phase observation. We used this method in the testing phase of our example as it is good for evaluating how a model might perform in real-time, as more data is collected and assimilated into the model fitting process. When using the *rolling-origin-update* window, we fit the model to the training data and forecast the testing phase sequentially as in the rolling-origin-recalibration window, however the testing data is only used as inputs for the model, which is not re-fit with the new testing data. This method is practical for evaluating models that are computationally intensive to fit, have specifically defined training-phase parameters that are of interest to the research, or have enough data that additional observations will not meaningfully affect the model fit.

However we decide to conduct our testing phase forecasts, we will compare the results using the evaluation metric that we specified in Section 1.3.2. While not absolutely necessary, we recommend building a simple benchmark model to which we can compare the results of our models. This benchmark model should make a reasonable guess at the outcome, which one would hope a trained forecasting model can outperform. For short-term forecasts in a setting with high autocorrelation (*e.g.* weekly dengue incidence), a good benchmark model could be an AR1 model that merely predicts the last observed value. For forecasts with longer horizons or less autocorrelation (*e.g.* annual dengue incidence), a good benchmark model could be a seasonal or long-term average. In some situations, there is an "industry standard" model that we should compare our novel methods to, such as SARIMA models for sequential forecasts across multiple horizons. Either using a relative metric or using a metric that can be compared to benchmark forecasts put the results from our models in context, which is helpful to officials, reviewers at journals, and ourselves. In our example, we evaluated the testing phase forecasts using relative mean absolute error (rMAE) of our model over baseline forecasts based on the ten-year median incidence

rate for each province. The univariate model had less testing phase error than the best cross-validation model (Figure 1.2) and had about 20% less error than the baseline forecasts (not shown).

When we are interpreting our results, we should focus on the model that performed best in the testing phase, especially when that model is not the best model from the training phase. If the best training phase model performs worse than a more parsimonious model in the testing phase, some feel a tendency to analyze the performance of the larger model during the training phase in their results. This is a reasonable reaction after the many long hours that were put into collecting, cleaning, and organizing dozens, hundreds, or even thousands of covariates that were either not selected or were subsequently outperformed by a measly 2-covariate forecasting model. A seemingly sound justification for such a practice is that the training phase is usually longer than that of the testing phase. However, our goal at the outset of this exercise was to find the model that makes the best forecasts that are generalizable outside of our data, which is defined by the performance on the testing phase. Specific times or places where a different model showed good results could be areas for future forecasting activities. Prior to splitting the data into training and testing, think about how many observations are needed for each sample; there need to be enough training observations to properly fit the model (depending on the model type) and there need to be enough testing observations to properly evaluate the model. With a short time-series, there may be too few data to split and thus only cross validation can be conducted on all of the data; in this scenario, interpretations about the model performance will be weaker than those with a separate testing phase.

1.4 Operationalizing forecasts for public health

As we have mentioned above, making forecasts on infectious disease data is difficult due to the culmination of a variety of factors, from the microscopic to the population

level, which are difficult to do in any circumstances, but are compounded by logistical issues when trying to make forecasts in real time. Logistical challenges include assimilating newly-collected data into the forecasting framework, accounting for delays in case reporting, and effectively communicating the model results to public health officials.

1.4.1 Reporting delays

Making forecasts in real time introduces the dimension of reporting delays into our forecasting models. From a disease surveillance perspective, reporting delays are a timeliness issue which varies by features of the disease (ease of diagnosis, incubation time), surveillance entity (local, state, or national government), transmission type (electronic or not), and case load, as well as variability in reporting between surveillance systems [90, 21]. From a data perspective, this means that observed values in the recent past are subject to change.

Since most forecasting models make the assumption that values used for fitting are fixed, we need to adapt our forecasting process by "nowcasting" observed values in the recent past. One method of now-casting is to only include "sufficiently complete" data up such that a forecasting model can make stable forecasts. For instance, 75% of dengue hemorrhagic fever cases in Thailand were reported to the Thai Ministry of Public Health within 10 weeks of infection [158]. To account for this, we ignored the last 12 weeks (actually 6 biweeks, to be exact) before forecasting forward. In our notation, we fit our model to data $y_{1:(t+k-1)}$ to make a k-step forecast, y_{t+k}, where $k = -6$. Another method of nowcasting is to use past reporting delays to model recent incomplete counts. Höhle and an der Heiden [85] provide a framework for nowcasting infectious disease incidence.

When case counts for prior time periods are subject to change, it is important for researchers to have a collection of data "snapshots", so that past situations can

be investigated retrospectively with the information that was available at the time. Thus, we should manage an independent database of cases as they are reported to us, containing date of illness and incidence that is timestamped upon deposit into our database.

1.4.2 Communication of results

Public health authorities have shown increasing interest in working with infectious disease forecasters in the light of recent important public health crises. Starting in 2009 with the pandemic influenza A outbreak, public health officials turned to forecasters for estimates of burden and burden averted due to vaccines and antivirals. During the Ebola outbreak in 2014, public health officials again turned to prediction for specific information regarding the potential outbreak size and intervention impacts. These efforts highlight how infectious disease forecasting can support public health practice now and in the future.

1.4.2.1 What makes a good forecast?

Previous work in meteorology has outlined 3 distinct forecast attributes of a forecast that contribute to its usefulness, or "goodness" [135]. If we apply these guidelines to infectious disease forecasting, we can surmise that a forecast is good if it is (a) *consistent*: reflecting the forecaster's best judgment, (b) *quality*: forecasts conditions that are actually observed during the time being forecasted, and (c) *valuable*: informs policy or other decision-making that results in increased benefits to individuals or society.

For a forecast to reflect the forecaster's "best judgment" means that the forecast is reasonable based on the forecaster's expert knowledge base, prior experience, and best and current methodology. The forecaster's internal judgments are not usually available for evaluation or quantification, but could say that a forecast is not a reflection of

best judgment if we discover that a forecasting model contains an error or under some conditions produces values outside the range of possible values.

To meet the conditions for high quality, forecasted values must correspond closely to observed values. The field of forecast verification is so vast and specialized that we could not possibly give it a comprehensive treatment here. Suffice it to say that reducing error is central goal of the field of forecasting. Examples of quality measurement approaches include the mean absolute error and the mean-squared error, which reflect forecast accuracy. Other examples include measures of bias, skill (often a comparison to reference models), and uncertainty [96].

Infectious disease forecasts are valuable if they are used to influence decisions. Sometimes value can sometimes be accessed in quantitative units (*e.g.* lives or money saved or lost). Forecast quality influences value to a large extent, but so do other more qualitative features of how the forecast is communicated. For example, a forecast will have a larger impact on decision-making if it is timely, presented clearly, and uses meaningful units in addition to being accurate or improving on a previous system.

1.5 Conclusion and Future Directions

There has been a great deal of progress made in infectious disease forecasting, however the field is very much still in its infancy. Forecasts of epidemics can inform public health response and decision-making, including risk communication to the general public, and timing and spatial targeting of interventions (*e.g.* vaccination campaigns or vector control measures). However, to maximize the impact that forecasts can have on the practice of public health, interdisciplinary teams must come together to tackle a variety of challenges, from the technological and statistical, to the biological and behavioral. To this end, the field of infectious disease forecasting should emphasize the development and integration of new theoretical frameworks that can be directly linked to tangible public health strategies.

To facilitate the development of scalable forecasting infrastructure and continued research on improving forecasting, the field should focus on developing data standards for both surveillance data and forecasts themselves. This will foster continued methodological development and facilitate scientific inquiry by enabling standard comparisons across forecasting efforts. One key barrier to entry to this field is that the problems are operationally complex: a model may be asked to forecast multiple targets at multiple different times, using only available data at a given time. Converging on standard language and terminology to describe these challenges is key to growing the field and will accelerate discovery and innovation for years to come.

CHAPTER 2

PROSPECTIVE FORECASTS OF ANNUAL DENGUE HEMORRHAGIC FEVER INCIDENCE IN THAILAND, 2010–2014

(The contents of this chapter are published in the Proceedings of the National Academy of Sciences of the United States of America[108], co-authored with Krzysztof Sakrejda, Evan L. Ray, Lindsay T. Keegan, Qifang Bi, Paphanij Suangtho, Soawapak Hinjoy, Sopon Iamsirithaworn, Suthanun Suthachana, Yongjua Laosiritaworn, Derek A.T. Cummings, Justin Lessler, and Nicholas G. Reich, and appears here with permission.)

Dengue hemorrhagic fever, a severe manifestation of dengue viral infection that can cause severe bleeding, organ impairment, and even death, affects between 15,000 and 105,000 people each year in Thailand. While all Thai provinces experience at least one DHF case most years, the distribution of cases shifts regionally from year to year. Accurately forecasting where DHF outbreaks occur prior to the dengue season could help public health officials prioritize public health activities. We develop statistical models that use biologically-plausible covariates, observed by April each year, to forecast the cumulative DHF incidence for the remainder of the year. We perform cross-validation during the training phase (2000-2009) to select the covariates for these models. A parsimonious model based on pre-season incidence outperforms the 10-year median for 65% of province-level annual forecasts, reduces the mean absolute error by 19%, and successfully forecasts outbreaks (AUC=0.84) over the testing period (2010-2014). We find that functions of past incidence contribute most strongly to model performance whereas the importance of environmental covariates

varies regionally. This work illustrates that accurate forecasts of dengue risk are possible in a policy-relevant time-frame.

2.1 Introduction

Dengue, a mosquito-borne virus prevalent throughout the tropics and sub-tropics, infects an estimated 390 million people every year [13]. While the majority of infections are mild or asymptomatic, the more severe forms of dengue infection – dengue shock syndrome (DSS) and dengue hemorrhagic fever (DHF) – can result in organ failure or death [161]. The number of symptomatic dengue infections has doubled every ten years since 1990, in contrast to the declining incidence of most other communicable diseases [185].

In Thailand, dengue infection is endemic with substantial annual and geographic variation in incidence across its 76 provinces and 13 health regions (Figure 2.1). Over the past 15 years, an average of 43,137 (range 14,952-106,320) DHF cases have been reported to the Thailand Ministry of Public Health (MOPH) each year. Within a typical year, incidence rates in different provinces can vary by an order of magnitude, with some provinces experiencing less than 10 DHF cases per 100,000 population and others over 100 per 100,000 population.

Public health officials must determine where to allocate resources to manage the problems caused by dengue viral infection. A newly-approved vaccine may be able to reduce the number of dengue infections, if properly regimented [53]. For those already infected, effective case management can reduce the case-fatality rate of severe dengue [98]. With sufficient advance notice, public health officials could implement prevention programs and conduct interventions in regions that have the highest epidemic risk. Effective long-term forecasts would provide more timely information to aid in prioritizing these public health activities.

Prior dengue forecasting efforts by members of our group and others have focused on short time scales (weeks or months) [207, 114, 83, 158, 94]. These studies demonstrated the importance of recent case counts and seasonality on the immediate trajectory of dengue incidence. In 2015, the National Oceanic and Atmospheric Administration (NOAA) and the Centers for Disease Control (CDC) hosted a competition to make within-season forecasts for annual dengue incidence, epidemic peak, and peak height for San Juan, Puerto Rico and Iquitos, Peru [138]. Groups that employed methods relying solely on functions of incidence performed well relative to baseline forecasts [209, 156] and were amongst the top performers in the competition [95].

Whether an infectious disease spreads within a population depends on the transmission rate of the disease and the number of susceptible individuals [102, 105], thus long-term forecasting models for DHF incidence may need to account for climatic factors that could affect transmission as well as population susceptibility. Climatic factors, such as temperature, rainfall, and humidity, may impact both the prevalence and distribution of the dengue vector, the *Aedes* mosquito [97, 174, 24], as well as the transmission efficiency of dengue virus [13, 92, 87]. During the low dengue season, these climatic factors may be indicative of incidence in the following high dengue season, perhaps due to their role in vector survival and larval development [30]. Even in ideal conditions for disease transmission, there needs to be a sufficiently large susceptible population for a disease to spread. Dengue has complex immunological dynamics that make tracking the number of susceptible individuals within a population difficult. The vast majority of first dengue infections are asymptomatic, while second infections are more likely to result in severe outcomes such as DHF and DSS [28, 48]. Infection by any of its four serotypes may offer temporary immunity to the other serotypes and lifelong immunity to the contracted serotype [161, 4, 202, 160], although there is some evidence that repeat infections of the same serotype may occur [56, 200].

A useful forecasting model needs to make better predictions than a baseline model on out-of-sample observations [159]. For decades, researchers have split their data into 'training' and 'testing' samples to separate the fitting and evaluation processes [187, 78]. Cross validation is a popular technique for estimating the expected prediction error, thus minimizing the cross-validation error on the training sample might be expected to improve predictions over the testing sample. However, this can lead forecasters to select models that "overfit" on the training sample and therefore do not perform well on the testing sample [141]. Hence, it is prudent for researchers to also select a parsimonious model with more cross-validation error that might perform better on out-of-sample data [78, 141]. In the testing phase, using a sensible baseline model as a comparison allows researchers to measure how much a forecasting model improves over a benchmark in an interpretable manner [88].

Using demographic, weather, and dengue data from 2000 to 2009, we selected two models using a cross-validated variable selection procedure to make probabilistic forecasts of the annual DHF incidence for 2010 to 2014. We chose to predict DHF cases because reporting for this severe form of dengue is thought to be more consistent across time and space, while still being a primary indicator of the burden of disease [158]. We compare the forecasts from these models to baseline forecasts derived from a province's median DHF incidence rate over the past ten years. We use the probabilistic distributions to estimate the outbreak risk for each province. We investigate features of our forecasting models, including regional variations in performance and the most informative covariates. In doing so, we show that producing accurate forecasts that add value for public health decision makers is a viable endeavor.

2.2 Results

2.2.1 Models selected for forecasting

We obtained data on DHF cases (from the MOPH), population (National Statistical Office of Thailand), and weather (NOAA) [158, 126, 127, 32, 50, 5]. These data were summarized across time frames ranging from one month to one year to create 34 covariates for consideration by our model selection algorithm (Tables 2.1 and S1). We calculated an additional covariate, 'estimated relative susceptibility', based on the assumption that an infected person will be protected against all dengue serotypes for a period of roughly two years [160]. We made forecasts using the data available in April of each year, the month when the MOPH has historically finalized the incidence reports obtained from all provinces for the prior calendar year. Hence, all "annual" forecasts are for DHF incidence between April and December of the year they are made. Across the 15 years used in this study, 87% of the DHF cases occurred between April and December of each year.

We used leave-one-year-out cross validation to predict the DHF incidence across the 760 province-years in the training phase (76 provinces for each year from 2000-2009). Of the 202 candidate models considered, the model with the smallest leave-one-year-out-cross-validated mean absolute error (CV MAE) included five covariates: pre-season (January-March) incidence rate, total January rainfall, mean January temperature, mean temperature during the low dengue season (November-March, henceforth 'low-season'), and population size (Figure 2.2). In order to avoid overfitting on the training phase, we also chose the model with the fewest covariates within one standard deviation of the minimum CV MAE [78]. Using this procedure, we selected a model that included only pre-season incidence. We refer to these models as the 'weather, incidence, and population (WIP) model' and 'incidence-only model', respectively.

Table 2.1. Justifications for types of covariates considered for inclusion prior to model selection

Covariate Type	Reason for inclusion
Incidence	Large dengue outbreaks may temporarily deplete the susceptible population [160, 202, 4]. Larger dengue seasons often start earlier [30].
Demographics	Higher population density may facilitate dengue transmission [191].
Humidity	Humidity may improve the survival rate of *Aedes* mosquito eggs [97, 30].
Rainfall	Rainfall is essential for *Aedes* mosquito breeding and may have a positive effect on dengue transmission [13, 174].
Temperature	Temperatures must be warm enough for *Aedes* mosquitoes to imbibe blood [24], but cool enough for optimal survival of eggs [97].

2.2.2 Forecasting performance in the testing phase

Across the 380 province-years in the testing phase (2010-2014), forecasts from the incidence-only model were more accurate than forecasts from the WIP model (relative mean absolute error [rMAE]=93% [88]) and baseline forecasts derived from the 10-year median incidence rate (rMAE=81%). The incidence-only model forecasts were closer to the observed DHF incidence than those of the WIP model in 217 of 380 (57%) province-years and better than baseline forecasts in 246 of 380 (65%) province-years (Table S2). In each year, the incidence-only model outperformed both the WIP model and the baseline forecasts in aggregate (*i.e.* the all-province MAE was lower and more forecasts were closer to the observed incidences) (Figure 2.3 and Table S3). Across all testing phase province-years, the 80% prediction interval from the incidence-only model covered 80% of the observed DHF incidences, compared to 70% covered by the WIP 80% prediction interval.

The testing-phase performance of each model varied across Thailand's 13 MOPH health regions (Figure S2). The incidence-only model performed best in 10 of 13 (77%) regions, the WIP model performed best in 2 of 13 (15%) regions, and the baseline forecasts performed best in 1 of 13 (8%) regions (Figure 2.4 and Table S4). The

WIP model made better forecasts, relative to the baseline forecasts, for regions that experience colder (MOPH regions 1, 7, and 8) or rainier (MOPH regions 11 and 12) low seasons than for the rest of Thailand. In these regions climatic suitability for mosquito breeding varies between years, hence a model with climate covariates can provide a strong early indication of annual incidence. Conversely, the WIP model performed especially poorly in Bangkok, which has consistently warm weather and moderate rainfall from year to year.

We quantified the risk of an outbreak for each province-year using samples from the predictive distributions of the incidence-only model. We define an 'outbreak' to be when a province experiences a DHF incidence rate that is greater than two standard deviations above its 10-year median rate. In the testing phase, there were outbreaks in 38 of 380 (10%) province-years. Across all testing phase province-years, the forecasted outbreak probability had a strong correspondence with the likelihood of a province experiencing an outbreak (Figure 2.5b). Correspondence was particularly good in the 360 province-years where forecasted outbreak probabilities were less than 0.5 (Figure 2.5a). Due to the unlikely nature of outbreaks, the incidence-only model only forecasted outbreak probabilities above 0.5 for 20 province-years (5% of all forecasts), however 8 of the 38 (21%) outbreaks occurred during these province-years. The incidence-only model correctly ordered the outbreak probabilities of any two randomly chosen province-years 84% of the time (Figure 2.5c) [77].

2.3 Discussion

We have shown that it is possible to make accurate forecasts of annual dengue hemorrhagic fever (DHF) incidence for Thailand at the province level using data available to policy makers prior to each year's dengue season. Testing forecasts from a parsimonious model performed better than forecasts based on 10-year median incidence rates. Further, this model successfully ordered provinces by their risk of experiencing

an outbreak. These forecasts can provide timely and valuable information to policy makers as they prepare for the coming dengue season. By integrating biological and statistical approaches, these models push the envelope on how early it may be possible to accurately forecast annual dengue incidence. However, further improvements are needed for these forecasts to have their maximum impact.

The inclusion of climatic covariates did not consistently add value to forecasts relative to the incidence-only model. While there is biological evidence that *Aedes* mosquitoes are affected by climatic factors [97, 13, 24], the usage of such factors in dengue forecasting efforts have shown mixed results [207, 114, 83, 94, 92, 174, 116, 117]. These findings suggest that the associations between climate covariates and dengue either differ across time and space or are spurious correlations. Alternatively, climate may be one of several necessary-but-insufficient factors, along with susceptibility and recent incidence, whose combination results in ideal conditions for dengue transmission. Building a forecasting model that incorporates interactions between covariates is an area for future work.

The relative estimated susceptibility covariate was not selected for inclusion in either of the final models. This crude approximation of a complex mechanistic feature of disease was a component of the best six-covariate model, however that model had a larger cross-validated mean absolute error during the training phase than the weather, incidence, and population (WIP) model. A susceptibility term built on our mechanistic understanding of the disease process that more accurately captures the transient cross-protection between dengue serotypes could add value to a forecasting model.

Although we have demonstrated our ability to successfully forecast DHF incidence prior to the dengue season, many of the planning activities of the Thailand Ministry of Public Health (MOPH) occur even further in advance, thus the ability to make forecasts earlier in the year may be useful for public health policy. Historically, the

MOPH has finalized each year's dengue reports in the following April. This effectively sets the earliest possible date annual forecasts can be made if they are to be based on complete data. An accurate model of reporting delays or more timely reporting could shift this date earlier. Likewise, forecasters could build a series of models optimized for data available at different times of the year.

To aid in the translation of this research into practice we created sortable spreadsheet reports with results for each year that were then disseminated within the MOPH. These reports are used for ranking provinces based on the forecasted probability of an outbreak and prioritizing locations for targeted interventions. This operational interpretation of the results emphasizes the importance of the relative rankings being accurate. The finding that 84% of the time our model would correctly rank two randomly-selected province-years by outbreak probability directly supports the use of these forecasts in practice.

Making timely forecasts of infectious disease incidence is a challenging but important task. Accurate forecasts could play an important role in implementing targeted interventions designed to reduce transmission, such as in helping to determine the location and timing of vector control activities and the mobilization of additional resources, as well as for reporting risk of infection to the public. Additionally, they could play a critical role in a systematic study of how well different interventions prevent or reduce the size of disease outbreaks. Collaborative efforts between public health agencies and academic- or industry-based teams with predictive modeling expertise are critical to helping propel this field forward. With the rapid growth and maturation of disease surveillance systems worldwide, developing our understanding of the best methods for creating and evaluating forecasts of infectious disease should continue to be a global health priority.

2.4 Materials and Methods

2.4.1 Weather covariate screening

To investigate the utility of weather for forecasting annual DHF incidence, we included a variety of temperature, humidity, and rainfall covariates across several seasonal periods (Table S1). We downloaded weather station data from the National Oceanic and Atmospheric Administration (NOAA), which provided daily rain and temperature estimates for weather stations in 35 provinces [126, 127]. Using the `stationaRy` [89] package in R [153], we obtained integrated surface data from the National Climatic Data Center (NCDC) [32]. These data consist of temperature and humidity measurements from weather stations in 65 provinces (including all 35 provinces from the NOAA dataset), at six-hour intervals. For all provinces, we downloaded monthly temperature and rainfall data on 0.5x0.5 latitude-longitude resolution from the Earth System Research Laboratory (ESRL) at NOAA [50, 5].

For the NOAA and NCDC weather station data, we found the most consistently reported weather station for each province and extracted the daily maximum and minimum temperature, maximum humidity, and rainfall. We aggregated these measures into monthly covariates for maximum, minimum, and mean temperature, maximum and mean humidity, and maximum and total rainfall across January, February, and March. We also aggregated weather covariates across the "low season", from November through March, when fewer DHF cases have occurred historically, on average. This time of season aligns with the dry season in Thailand, which has reduced temperatures and precipitation as compared to the high dengue season, from April through October, which corresponds with the rainy season.

We removed any covariates for which more than half of the aggregated observations from one source were missing. For example with NOAA data, if 263 province-years (half of 35 provinces for 15 years) of observations were missing for a covariate, it was removed; as was the case for low-season minimum and maximum temperature. The

ESRL data, from which the three covariates in the WIP model were derived, had one observation per month and was completely reported across all provinces.

2.4.2 Relative estimated susceptibility

The estimated relative susceptibility covariate is a standardized rolling sum of cases from the previous two years. This is based on the approximate duration of time after infection with one dengue serotype that an individual may experience cross-protection to a subsequent heterologous infection [160]. We calculate this quantity with the following equations:

$$s_{i,t} = s_{i,t-1} - \frac{y_{i,t-1}}{n_{i,t-1}} + \frac{y_{i,t-3}}{n_{i,t-3}}$$
$$s_{i,0} = \frac{1}{10} \sum_{t=2000}^{2009} \frac{y_{i,t}}{n_{i,t}},$$

where $s_{i,t}$ is the estimated relative susceptibility, $y_{i,t}$ is the observed incidence, and $n_{i,t}$ is the population in province i in year t. Each year, the susceptibility for the prior year ($s_{i,t-1}$) is updated by removing the people who were infected in the past year ($\frac{y_{i,t-1}}{n_{i,t-1}}$), as we assume that they are immune to one serotype of dengue and cross-protected against the other serotypes. Furthermore, the cross-protection for people who were infected three years prior ($\frac{y_{i,t-3}}{n_{i,t-3}}$) will have worn off and they are reintroduced to the pool of susceptibles. We assume that each province starts with an estimated relative susceptibility equal to the average incidence rate over the training phase ($s_{i,0}$). This accounts for the fact that provinces with larger susceptible populations are more likely to have greater incidence than provinces with smaller susceptible populations [102]. When there is no data for the year three years prior, $s_{i,0}$ is used in place of $\frac{y_{i,t-3}}{n_{i,t-3}}$. Using rates instead of raw counts yields a covariate that can be compared across provinces with different population sizes. Though there are more cases of non-hemorrhagic

dengue fever and asymptomatic cases than observed DHF cases, DHF cases may serve as a proxy for the underlying disease dynamics [13].

2.4.3 Model structure and estimation

The model that we used to forecast annual DHF incidence for this study is a generalized additive model [78]. Specifically, we use a generalized additive model with a negative binomial family, separate penalized smoothing splines for each covariate, and province-level random effects:

$$Y_{i,t} \sim \text{NB}(n_{i,t}\lambda_{i,t}, r), \tag{2.1}$$

$$\log\Big[\mathbf{E}(Y_{i,t})\Big] = \beta_0 + \log(n_{i,t}) + \alpha_i + \sum_{j=1}^{J} g_j(x_{j,i,t}|\boldsymbol{\theta}), \tag{2.2}$$

$$\alpha_i \sim \text{Normal}(\mu, \sigma^2). \tag{2.3}$$

We model the incidence ($Y_{i,t}$) for province i in year t as following a negative binomial distribution with the mean equal to the province population ($n_{i,t}$) times the incidence rate ($\lambda_{i,t}$) and a dispersion parameter r. After a log transformation, we model the mean of this distribution using an intercept (β_0), a random effect for each province (α_i) and a cubic spline for each of J covariates ($g_j(x_{j,i,t}|\boldsymbol{\theta})$).

To obtain predictive distribution samples, we use a two-stage procedure to incorporate the uncertainty from our model parameter estimates and from the negative binomial distribution. We first draw 100 sample parameter sets from a multivariate normal distribution with mean equal to the point estimates of the parameters ($\boldsymbol{\theta}, \mu, \sigma^2$) from Equations (2.2)-(2.3) and covariance equal to the matrix of standard errors. Each of these sampled parameter sets yields a corresponding $\widehat{\lambda}_{i,t}$. We then draw 100 samples from the negative binomial distribution given in Equation (2.1) for each $\widehat{\lambda}_{i,t}$ with the fixed estimate of r to obtain a sample of size 10,000 from the predictive distribution

for $Y_{i,t}$. We calculate the point estimate for each province-year, $\widehat{Y}_{i,t}$, as the median of these samples from the predictive distribution. The lower and upper limits of the 80% prediction intervals were defined by taking the 10^{th} and 90^{th} percentiles of these samples from the predictive distribution.

2.4.4 Model selection algorithm

To choose the covariates to include in the forecasting models, we used a forward-backward stepwise algorithm to minimize the leave-one-year-out-cross-validated mean absolute error (CV MAE) during the training phase [42]. Starting with a null model, we iteratively added or removed the covariate that reduced the CV MAE the most at each step. The model with the smallest CV MAE at the end of the iterative process was the WIP model. To guard against the possibility of overfitting, we also selected the nested model with the fewest covariates within one standard deviation of the WIP model CV MAE [78], which was the incidence-only model.

In order to choose the number of knots for each covariate spline, we cross-validated every single-covariate model varying the number of knots from 3 to 8, which we conducted prior to the forward-backward stepwise algorithm above. We chose the model with the fewest knots within one standard deviation of the smallest CV MAE for each covariate. We fixed this number of knots for each covariate spline for all multivariate models.

2.4.5 Mean absolute error

We used mean absolute error (MAE) as our metric to select models during the training phase and relative mean absolute error (rMAE) to evaluate the models during the testing phase. Forecasts were made on the log scale, thus our MAE took the form:

$$\text{MAE} = \frac{1}{P_k} \sum_{i,t \in k} \left| \log(\widehat{Y}_{i,t}) - \log(Y_{i,t}) \right| = \frac{1}{P_k} \sum_{i,t \in k} \left| \log \left(\frac{\widehat{Y}_{i,t}}{Y_{i,t}} \right) \right|$$

where P_k is the total number of province-years in block k, which could be the entire training or testing phase, or subset to one year, province, or region. This form of the MAE has the interpretation that precision is relative to magnitude; *e.g.* predicting an incidence of 12 when an incidence of 7 is observed would have the same absolute error as predicting an incidence of 120 when an incidence of 70 is observed ($\log(\frac{12}{7}) = \log(\frac{120}{70}) = 0.539$).

The testing phase point predictions were compared to baseline forecasts using rMAE, an intuitive, scalable, and stable metric for evaluating forecasts [159]:

$$\text{rMAE} = \frac{\text{MAE}_{\text{model}}}{\text{MAE}_{\text{baseline}}}.$$

This metric can be interpreted as the percentage of error observed in the forecasting model relative to that in the baseline forecasts; *e.g.* if $\text{MAE}_{\text{model}} = 0.6$ and $\text{MAE}_{\text{baseline}} = 0.8$, then the forecasting model's predictions were 25% closer to the observed value than the baseline forecasts.

2.4.6 Data and code availability

All data processing and analysis was performed in R version 3.3.1 (2017-03-16) [153]. The code and data for this analysis is publicly available at `https://doi.org/10.5281/zenodo.814994`.

2.5 Acknowledgements

This project was funded by NIH NIAID grant 1R01AI102939 and NIGMS grant R35GM119582. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the National Institutes of Health or the National Institute of General Medical Sciences. The funders had no role in study design, data collection and analysis, decision to present, or preparation of the presentation.

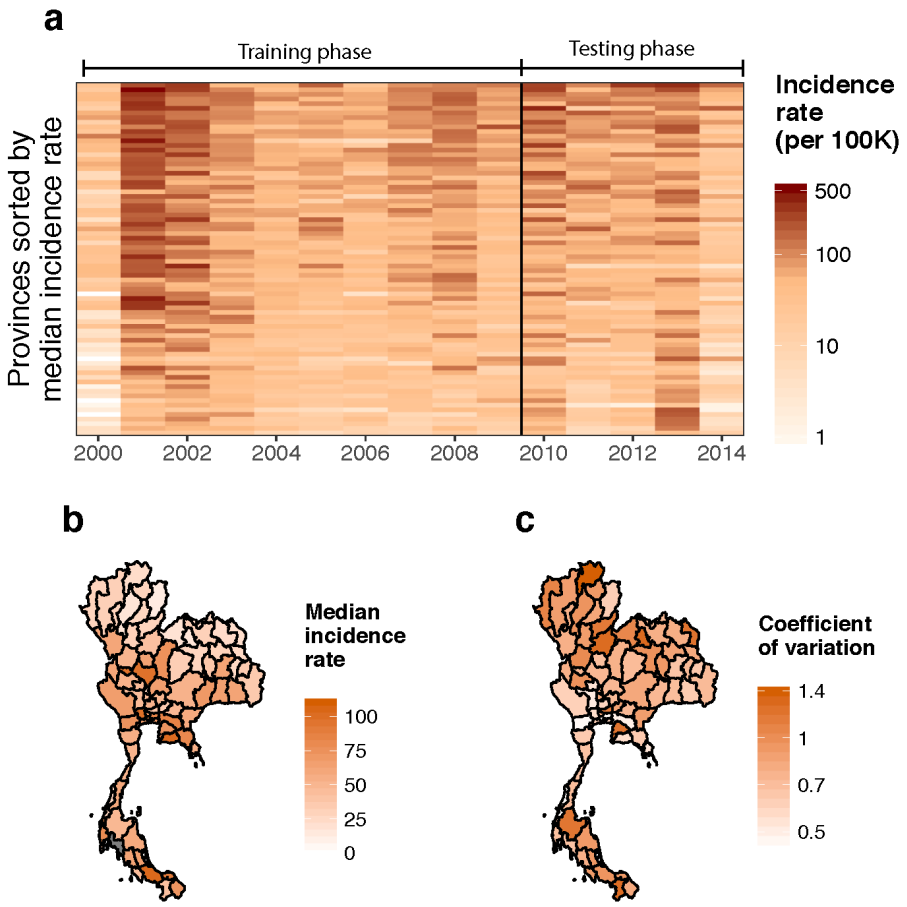

Figure 2.1. The temporal and spatial distribution of annual dengue hemorrhagic fever (DHF) incidence rates in Thailand. **(a)** The annual DHF incidence rate, per 100,000 population, for each Thai province and year used in this study. **(b)** The median annual DHF incidence rate, per 100,000 population, for each province from 2000-2014. **(c)** The coefficient of variation (standard deviation divided by the mean) of the annual DHF incidence rate for each province.

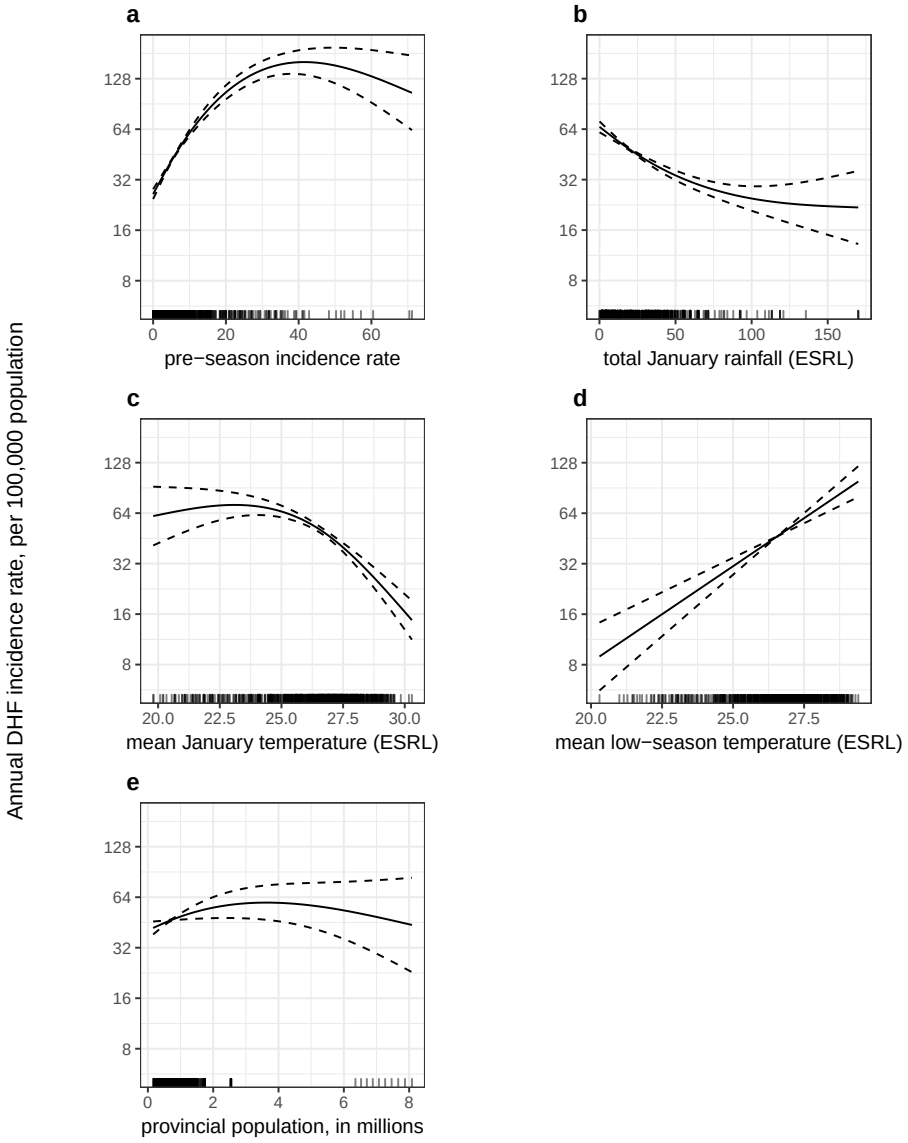

Figure 2.2. Weather, incidence, and population (WIP) model covariate fit curves. The solid lines represent the average association between each covariate in the WIP model and annual dengue hemorrhagic fever (DHF) incidence per 100,000 population during the training phase, fixing all other covariates at their mean. The dashed lines are the confidence intervals of each association, defined as two standard errors above and below the mean association. The covariates are arranged by performance in the Wald test from largest reduction in deviance **(a)** to smallest reduction in deviance **(e)**.

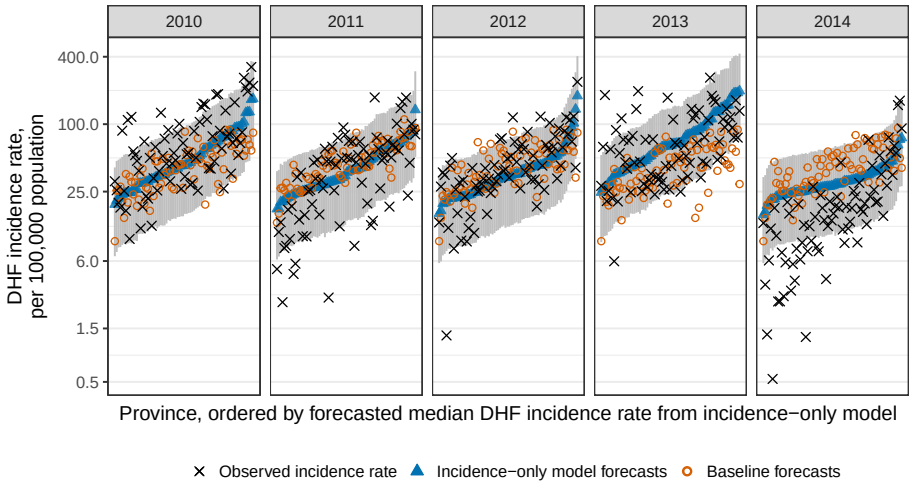

Figure 2.3. Incidence-only model forecasts for each year of the testing phase compared to the baseline forecasts and the observed values. Forecasts for the annual dengue hemorrhagic fever (DHF) incidence rate, per 100,000 population, from the incidence-only model (blue triangles with gray 80% prediction intervals), baseline forecasts (red circles), and observed values (black x's) for each province and year in the testing phase.

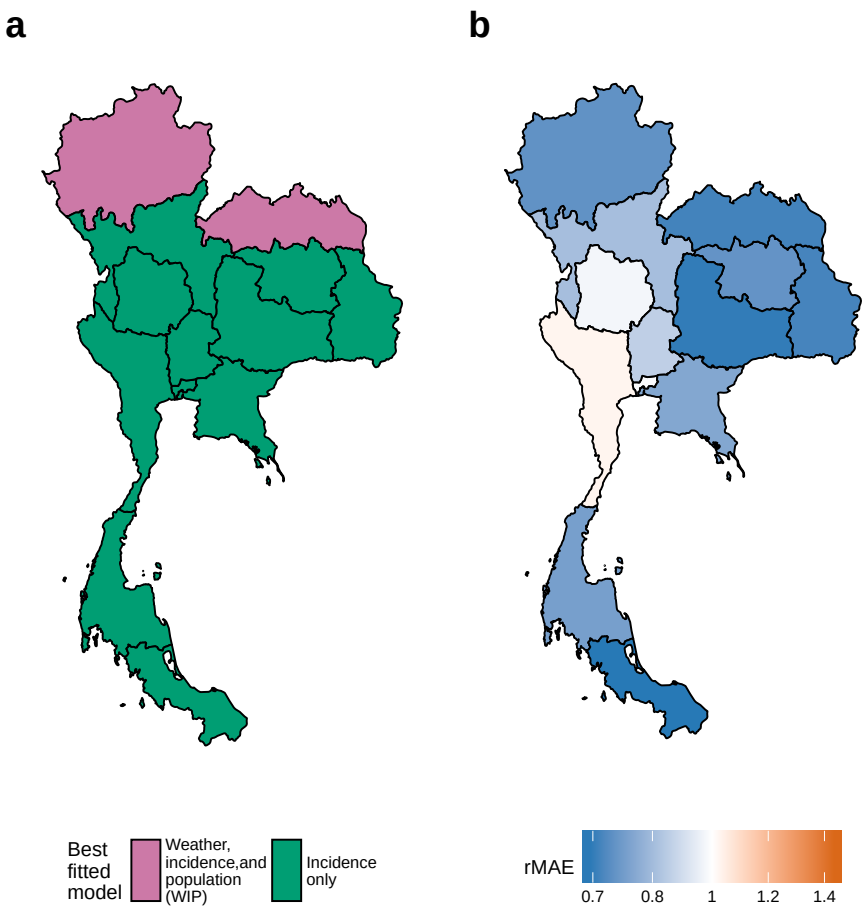

Figure 2.4. Geographic variation in model and performance. **(a)** The best fitted model in the testing phase for each Ministry of Public Health (MOPH) region, which shows spatial patterns of performance. **(b)** The relative mean absolute error (rMAE) of the forecasts for each MOPH region from the models in (a) over the baseline forecasts, *i.e.* the two northernmost MOPH regions show the rMAE of the WIP model forecasts, while the rest show the rMAE of the incidence-only model forecasts. Areas with: less error than the baseline are blue, more error than the baseline are red, and equal to the baseline are white.

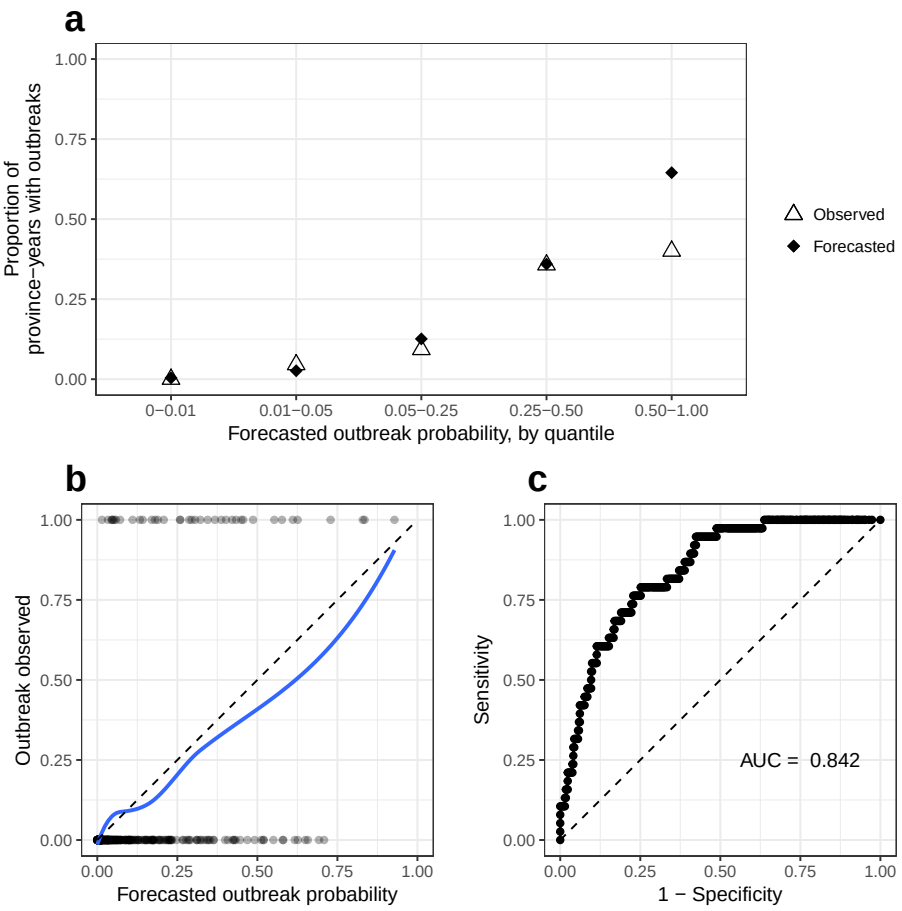

Figure 2.5. The performance of outbreak forecasts by the incidence-only model. **(a)** The proportion of province-years that observed an outbreak by their forecasted outbreak probability, which are binned into quantiles. An outbreak is defined as an annual dengue hemorrhagic fever (DHF) incidence rate greater than two standard deviations above the median annual DHF incidence rate for the past ten years. For each forecasted outbreak quantile, the black diamonds indicate the expected proportion of province-years with an outbreak based on incidence-only model forecasts and the hollow triangles indicate the observed proportion of province-years with an outbreak. **(b)** The forecasted probability of an outbreak for each province-year in the testing phase and whether or not an outbreak was observed. The blue loess smoothed line shows the probability of observing an outbreak for a given forecasted outbreak probability from the incidence-only model. **(c)** The receiver operating characteristic (ROC) curve based on the incidence-only model's sensitivity and specificity on outbreak forecasts. The area under the ROC curve (AUC) is indicated below the line of no-discrimination (dashed).

CHAPTER 3

THE COVARIATE-ADJUSTED RESIDUALS ESTIMATOR AND ITS USE IN BOTH RANDOMIZED TRIALS AND OBSERVATIONAL SETTINGS

3.1 Introduction

Estimating the effect of an exposure on a population has a long history in observational epidemiology. In 1855, John Snow compared the mortality rates of households in London by the company that supplied their water to locate the source of a cholera epidemic.[183] In 1881, Louis Pasteur inoculated 50 sheep with anthrax, 25 of whom had been vaccinated; the vaccinated sheep survived as the unvaccinated died, proving that his anthrax vaccine was effective.[128] In 1948, Austin Bradford Hill conducted the first modern randomized clinical trial, to evaluate a treatment for pulmonary tuberculosis,[122, 214] and later formulated guidelines for researchers and practitioners to transition from statistical association to causation.[84] Since then, there has been a proliferation of methods to determine the exposure effects in randomized trials and observational studies.[55, 33, 34, 192, 131, 168, 9]

One such method is the covariate-adjusted residuals estimator (CARE), which was formulated to estimate the effect of an exposure on an outcome of interest in individually randomized or cluster-randomized trials.[60, 10, 79] To implement CARE, researchers first predict the outcome of interest using baseline covariates that influence the outcome, while leaving out the exposure. Then they find the average prediction error for each group; this error can be the difference between or the ratio of the predicted and observed values (*i.e.* the residuals). The CARE estimate of the exposure effect is the discrepancy between the average residuals in each group; this

discrepancy can be a difference, ratio, or other measure. Gail *et al.* demonstrated that CARE could increase the statistical power over an unadjusted estimator, while maintaining confidence interval coverage, by using parametric regression models to adjust for covariates that predicted the outcome in randomized trials.[60] Bennett *et al.* showed that CARE made estimates that were robust to small sample sizes and moderate imbalances in the distribution of predictive covariates.[10] To the best of our knowledge, the use of CARE with non-parametric predicted outcomes has not been evaluated.

CARE is commonly used in ecology under the name 'residual index',[91, 107, 19, 47, 171, 35] although it has received some criticism. In the ecological field of allometry, researchers have used the residual index to estimate the effect of an exposure on the body mass of an organism, often in observational settings rather than randomized trials. While there are domain-specific questions about whether ordinary least squares linear regression is being used appropriately in allometry,[70] others have questioned the statistical validity of the residual index under any circumstances. Garcia-Berthou stated that "even if the assumptions of the linear model hold for the original variables, they will not hold for the residuals" and thus "the 'residual index' should never be used for statistical analyses of condition or any other variable".[61] To the best of our knowledge, there has been no statistical theory presented to date to support the continued use of the residual index, and thus CARE, in an observational setting.

In this manuscript, we provide new non-parametric theory that shows CARE is a consistent estimator of the exposure effect in both randomized trials and observational settings and which assumptions are necessary for that to hold. Our work supplements and generalizes existing parametric results from Gail *et al.* for randomized trials to observational settings. We compare CARE to existing estimators and introduce a novel estimator for use in both randomized and observational settings that joins CARE with methods using inverse probability of treatment weighting.[164, 167] We

support our theory with two simulation studies and an application in an infectious disease setting.

As an illustration, we estimate the effect of bednets on childhood mortality in a cluster-randomized trial in Ghana as originally published by Binka *et al.*.[10, 79, 16] In this trial, the Kassena-Nankana region of Ghana was divided into 96 clusters, 48 of which were randomly selected to receive impregnated bednets in June 1993. From July 1993 to June 1995, children aged 6-59 months were surveilled until they died (the outcome of interest), they migrated out of the study area, they turned 60 months of age, or the end of the follow-up period was reached. The clusters had 138 to 439 children each, with an average of 274.4 children. The data includes age in months at time of enrollment, sex, outcome, person-years of follow-up, and the cluster-level exposure assignment for each child. To improve the precision of their analysis, Binka *et al.* used covariate-adjusted residuals to control for the imbalanced age distributions between the exposure levels. The authors found that bednets reduced all-cause child mortality by 17%. Hayes and Moulton extended this analysis and found that covariate-adjusted residuals produced a stronger effect of bednets on childhood mortality with less variance than the unadjusted estimator, both when using the relative rate and absolute difference.[79] We will use this case study as an example when describing the causal framework and statistical theory, as the basis for one of the simulation studies, and as our real data application.

3.2 Causal framework

Did bednets reduce childhood mortality in Ghana? This is a common causal scientific question: how would a change in an exposure (*e.g.* bednets) change an outcome (*e.g.* childhood mortality). As a result, answering causal questions require a different approach than descriptive or predictive questions. For example, a descriptive analysis may provide point and uncertainty estimates for the childhood mortality in

clusters that actually received bednets and in clusters that did not receive bednets. If we were interested in predicting childhood mortality, we would want to know whether including a covariate for bednets added value to predictions that may use other information (such as age and sex), regardless of whether that relationship was causal or associative. Answering the causal question requires a deeper understanding of the system that generates the exposure and the outcome, and encoding that understanding in the model used for analysis.

A causal framework is an infrastructure that can guide our answering of causally-motivated questions. The key components of a causal framework are the current knowledge of the data-generating process (represented by a *causal model*); the quantity that answers the scientific question (the *causal parameter*); the assumptions required to link the causal quantity to a well-defined function of the observed data distribution (the *statistical parameter* which may or may not be *identifiable*); and the estimation and inference of the statistical parameter.[147, 148, 149, 151]

A causal model is a structural framework for expressing the relationships between variables in a given setting.[147, 148, 149, 67, 46] A causal model can be expressed graphically as a diagram, where variables are connected by edges (arrows) that originate at a potential cause and terminate at the effect. Figure 3.1a is a diagram representing a randomized trial, like that of our case study, where A is a binary exposure ($A = 1$ if the cluster received bednets, $A = 0$ if the cluster did not receive bednets) and W^Y is the set of baseline covariates (the average age and percentage of children who are female for each cluster) that influence the outcome Y (childhood mortality). There are no edges pointing to the exposure A because the randomization procedure makes the allocation of exposure independent of all other covariates. For this experimental setting, we assume that this causal model describes the data generating process for each cluster and that clusters are causally independent (*i.e.* the outcome for one cluster

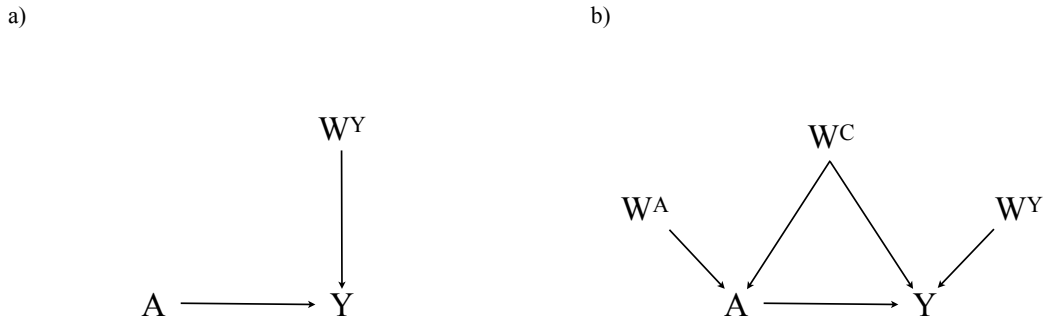

Figure 3.1. Causal diagrams for randomized trials **(a)** and observational studies **(b)**. These diagrams give a visual representation of the relationships between the variables in a causal model. Arrows are drawn from a potential cause to an effect. In a completely randomized trial setting, the exposure A is generated independently of all other variables and the outcome Y is influenced by both the exposure A and a set of baseline covariates W^Y. In an observational setting, the exposure A is no longer randomized, but instead is influenced by baseline covariates. Some of these covariates W^C also influence Y, thus confounding the relationship between the exposure A and the outcome Y. Other covariates W^A only influence the exposure A and not the outcome Y; as before some covariates W^Y only influence the outcome Y and not the exposure A. (For simplicity, other unmeasured sources of variation are omitted; see Appendix B.3.1 for a complete graph).

is only influenced by that cluster's exposure and baseline covariates and independent of the exposures, baseline covariates, and outcomes of the other clusters).

In an observational setting (as portrayed in Figure 3.1b), the allocation to the exposure A is not randomized and is potentially influenced by the baseline covariates. In addition to the covariates W^Y that influence the outcome Y, but not the exposure A, there are two new subsets of covariates. One subset of covariates W^A only influence the exposure A, but not the outcome Y. The other subset are confounding covariates W^C that influence both the exposure A and the outcome Y and thus obscure the isolation of the causal effect of interest. As a running example, we consider a scenario where bednets are distributed to clusters by the determination of local health officials instead of at random. In this scenario, consider a new baseline covariate: prior childhood mortality rate. Places with more childhood mortality prior to the intervention may be high risk for future childhood mortality and public health officials would want to concentrate their efforts in these areas. Thus prior childhood mortality rate is a confounding covariate, as it is a common cause of both the outcome and the exposure.

With the causal model specified, we can translate our scientific question into a causal parameter. We assume that the relationships within the causal model are autonomous, meaning that changing one relationship does not change the other relationships, though changes to causes may result in different effects downstream.[149] Thus, we could intervene to give impregnated bednets ($A = 1$) to all of the clusters in our target population to generate the counterfactual (hypothetical) outcome $Y(1)$, leaving the other relationships the same. Likewise, we could intervene to put all of the clusters in the unexposed group ($A = 0$) to find the counterfactual (hypothetical) outcome $Y(0)$, leaving the other relationships the same. With these counterfactual outcomes, we translate this scientific question into a well-defined causal quantity, specifically the average treatment effect (ATE):

$$ATE = \mathbb{E}[Y(1) - Y(0)]. \tag{3.1}$$

This is the difference in the average childhood mortality rate if all of the clusters in our target population received impregnated bednets ($Y(1)$) and if none of the clusters received impregnated bednets ($Y(0)$). We cannot directly measure this parameter because we only observe the outcomes Y corresponding to the actual exposures A and not both counterfactual outcomes. Thus, we need to outline the conditions and assumptions necessary to identify the causal parameter using a statistical parameter based on the data.

In our application, we use the difference in conditional expectations between the exposed and unexposed, adjusted for and averaged across the measured confounding covariates, as the statistical parameter:[1]

$$\Psi = \mathbb{E}_{W^C}\big[\mathbb{E}(Y \mid A = 1, W^C) - \mathbb{E}(Y \mid A = 0, W^C)\big]. \tag{3.2}$$

This statistical parameter Ψ identifies the ATE under two assumptions. First, there must be no unmeasured confounding between the exposure and the outcome, $Y(a) \perp\!\!\!\perp A \mid W^C$. Second, the 'positivity assumption', which states that each strata of measured confounding covariates have a non-zero probability of assignment to each exposure group ($\mathbb{P}(A = a|W^C = w^C) > 0, \forall w^C \in \mathbb{P}(W^C = w^C) > 0$), must hold.[150] These assumptions are satisfied differently between randomized and observational settings. In our example, the statistical parameter Ψ is the difference in the expected childhood mortality among clusters with the same common causes, with and without bednets, standardized with respect to the distribution of the confounding covariates in the population.

[1]This equation is known as the "G-computation identifiability result" in causal inference.[163]

In a randomized trial, the assumptions of no unmeasured confounding and positivity are often satisfied naturally by the study design. If the process for allocating units to each exposure level is truly random (*i.e.* a coin flip), then the exposure A is independent of the outcome Y and the baseline covariates W^Y, and thus there are no confounding covariates. This is why the assumption of no unmeasured confounding is also known as the 'randomization assumption'. Without confounding covariates, all units are equally likely to receive the exposure and the positivity assumption simplifies to $0 < \mathbb{P}(A = 1) < 1$ (for binary exposures). Since these assumptions hold by design, we can identify the ATE with the target statistical parameter Ψ^{RCT}, the difference in the conditional expectation between exposure groups.

$$\Psi^{RCT} = \mathbb{E}(Y|A = 1) - \mathbb{E}(Y|A = 0). \tag{3.3}$$

In a randomized trial where we observe outcome Y_i for $i = 1, \ldots, n$ units with exposure level A_i, the statistical parameter Ψ^{RCT} can be consistently estimated using the difference in the average outcome between the exposure groups, also known as the 'unadjusted estimator':[140]

$$\hat{\Psi}^{unadj} = \frac{1}{n_1} \sum_{i \forall A_i = 1} Y_i - \frac{1}{n_0} \sum_{i \forall A_i = 0} Y_i, \tag{3.4}$$

where n_a is the number of units in exposure level $A = a$. The proof showing that the unadjusted estimator is consistent for the statistical parameter Ψ^{RCT} in randomized trials is in the Supplementary Materials B.3.2.

In observational settings, the study design alone does not protect against confounding covariates or violations of the positivity assumption. The statistical parameter that we used for randomized trials Ψ^{RCT} (3.3) no longer identifies the ATE, therefore we must use the more general statistical parameter Ψ (3.2). In our running example, prior childhood mortality both influences present childhood mortality and is used by

public health officials to determine which clusters receive bednets. The unadjusted estimator does not account for this common cause and thus would be biased for the target statistical parameter Ψ (proof is in Supplementary Materials B.3.3). If we measure all of the common causes of the exposure and the outcome W^C (*i.e.* there is no unmeasured confounding) then the inverse probability of treatment weighting (IPTW) estimator can be used to estimate the statistical parameter Ψ:[164]

$$\hat{\Psi}^{IPTW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{1 - \hat{\mathbb{P}}(A_i = 1 \mid W_i^C)}\right) Y_i, \tag{3.5}$$

which controls for the confounding covariates through estimates of the probability of receiving the exposure, called 'propensity scores' $\hat{\mathbb{P}}(A = 1 \mid W^C)$.[167] If the propensity scores are consistent for the true conditional probability of exposure given common causes $\mathbb{P}(A = 1 \mid W^C)$, the IPTW estimator is consistent for the statistical parameter Ψ (proof in Supplementary Materials B.3.4). Notably, the propensity scores do not need to account for other covariates that influence the exposure W^A.

Positivity violations can be particularly problematic for the IPTW estimator. In observational settings, propensity scores close to zero or one can lead to highly variable estimates.[150] Therefore, accounting for W^A is not only unnecessary, but potentially harmful if it leads to extreme propensity scores. In randomized trials, the IPTW estimator can account for imbalances in the distributions of covariates that influence the outcome W^Y between exposure levels, leading to efficiency gains over the unadjusted estimator.[9, 196, 180]

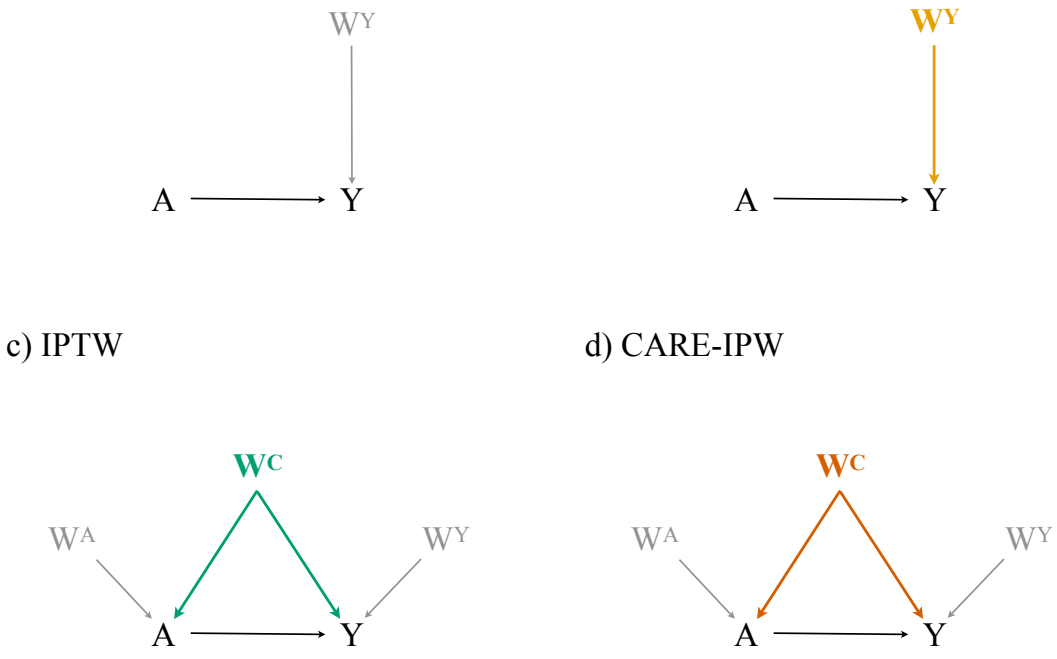

Figure 3.2. Diagrams that indicate which covariates are used by each estimator. **(a)** The unadjusted estimator does not use any covariates and only compares the average outcome Y between exposure levels $A = 0$ and $A = 1$. The unadjusted estimator is consistent for the target statistical parameter in randomized settings Ψ^{RCT} (shown), but not observational settings. **(b)** The covariate-adjusted residuals estimator (CARE) incorporates baseline covariates to make predictions for the outcome $\hat{\mathbb{E}}(Y \mid W^Y = w^Y)$. If these baseline covariates are predictive of the outcome and imbalanced between exposure levels, then CARE should be more efficient than the unadjusted estimator in randomized settings. CARE is not consistent for Ψ in observational settings with an exposure effect. **(c)** The inverse probability of treatment weighting (IPTW) estimator is consistent for Ψ when its propensity scores $\hat{\mathbb{P}}(A = 1 \mid W^C = w^C)$ are consistently estimated in observational settings. **(d)** When CARE is augmented by inverse probability weighting (CARE–IPW), it is consistent for Ψ when its propensity scores $\hat{\mathbb{P}}(A = 1 \mid W^C = w^C)$ are consistently estimated. CARE–IPW may be more efficient than the IPTW estimator when accounting for W^C in its predictions of the outcome $\hat{\mathbb{E}}(Y \mid W^C = w^C)$.

3.3 The covariate-adjusted residuals estimator (CARE)

3.3.1 CARE in randomized trials

The covariate-adjusted residuals estimator (CARE) was proposed as a method to estimate the coefficient for the exposure term in a parametric regression for the outcome Y, assuming no interactions between the exposure A and baseline covariates W^Y, in randomized trials.[60, 10, 79] With CARE, the outcome Y is predicted using only the baseline covariates W^Y and not the exposure A, giving us predicted values $\hat{\mathbb{E}}(Y \mid W^Y)$. Residuals are derived for each level of exposure; these residuals are commonly the difference between or the ratio of the predicted and observed values. The discrepancy in the average residuals between exposure levels gives the point estimate:

$$\hat{\Psi}^{CARE} = \underbrace{\frac{1}{n_1} \sum_{i \forall A_i = 1} \left[Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^Y) \right]}_{\text{Average residual for exposed}} - \underbrace{\frac{1}{n_0} \sum_{i \forall A_i = 0} \left[Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^Y) \right]}_{\text{Average residual for unexposed}} \tag{3.6}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)} \right) \left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C) \right), \tag{3.7}$$

where the number of units at each exposure level is equal to the total number of units times the empirical probability of exposure $n_a = n \times \hat{\mathbb{P}}(A = a)$. To obtain the predictions of the outcome $\hat{\mathbb{E}}(Y_i \mid W_i^C)$, Hayes and Moulton recommend using Poisson regression for event rates, logistic regression for binary outcomes, and linear regression for continuous outcomes.[79]

To the best of our knowledge, we are the first to non-parametrically prove that CARE provides a consistent estimator of the statistical parameter Ψ (3.2) and thus the ATE in a randomized trial, where the identifiability assumptions hold by design (Appendix B.1.1). CARE can be rearranged as the difference between the unadjusted estimator (3.4) and a second term incorporating the predictions of the outcome:

$$\text{Predicted} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)}\right)\hat{\mathbb{E}}(Y_i \mid W_i^C). \tag{3.8}$$

As discussed in Section 3.2, the unadjusted estimator is consistent for the target statistical parameter in randomized trials Ψ^{RCT} (3.3).[140] The 'predicted' term converges to zero, as the estimates of the outcome do not include the exposure A and the distributions of W^Y for each group are asymptotically equivalent in a randomized trial. In finite samples, CARE is expected to be a more efficient estimator than the unadjusted estimator when accounting for predictive covariates that may be imbalanced between the two groups in a randomized trial due to chance.[55, 33, 34, 192, 131, 168, 9]

When the predicted values of the outcome $\hat{\mathbb{E}}(Y \mid W^Y)$ are a constant value (*e.g.* zero or the mean of all observations $\bar{Y}$) then CARE is equivalent to the unadjusted estimator. This is plain to see in (3.8), as the average of a constant in the exposed group is equal to the average of the same constant in the unexposed group. Thus, the unadjusted estimator could be considered a special case of CARE.

3.3.2 CARE in observational studies

In observational studies with an exposure effect, CARE is not consistent for the statistical parameter Ψ (3.2) and thus will not provide an estimate of the ATE, even when the identifiability assumptions hold (Appendix B.1.2). The unadjusted component of CARE is not consistent for Ψ because the empirical probability of exposure $\hat{\mathbb{P}}(A = 1)$ is not consistent for the true conditional probability of exposure given confounding covariates $\mathbb{P}(A = 1 \mid W^C)$. For the same reason, the predicted component does not converge to zero. Therefore, CARE is not consistent for the statistical parameter Ψ in observational studies where there are confounding covariates W^C and a non-zero effect of exposure.

Under the strong null hypothesis that there is no exposure effect for all units, CARE is consistent for the target statistical parameter Ψ if the predicted values converge

to the true conditional mean outcome $\hat{\mathbb{E}}(Y \mid W^C) \rightarrow \mathbb{E}(Y \mid W^C, A) = \mathbb{E}(Y \mid W^C)$. However, since we do not know *a priori* whether or not the null hypothesis is true (which is presumably why we are trying to estimate the exposure effect), we do not recommend for CARE to be used in observational settings.

3.3.3 Improving upon CARE with inverse probability of treatment weighting

The inverse probability of treatment weighting (IPTW) estimator controls for measured confounders by upweighting outcomes that have a rare exposure-covariate combination (relative to a randomized trial) and downweighting those with a common exposure-covariate combination (again relative to a randomized trial) using propensity scores $\hat{\mathbb{P}}(A = 1 \mid W^C)$. As noted in Section 3.2, the IPTW estimator is consistent for the statistical parameter Ψ (3.2) when the propensity scores are consistent for the true conditional probability of exposure given confounding covariates $\mathbb{P}(A = 1 \mid W^C)$. This suggests that replacing the empirical probabilities of exposure $\hat{\mathbb{P}}(A = 1)$ with propensity scores $\hat{\mathbb{P}}(A = 1 \mid W^C)$ as an improvement to CARE (3.7):

$$\hat{\Psi}^{CARE-IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{1 - \hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} \right) \left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C) \right). \tag{3.9}$$

To our knowledge, this is estimator has not been previously proposed or explored.

As shown in Appendix B.1.3, when the propensity scores are consistently estimated, CARE–IPW is consistent for the target statistical parameter Ψ^{RCT} in randomized trials (3.3) and Ψ in observational settings (3.2). As with CARE, we can split CARE–IPW into two components and evaluate their expectations independently. The first component of CARE–IPW is equivalent to the IPTW estimator (3.5), which is consistent for the target statistical parameter Ψ when the propensity scores are

consistently estimated. The second component incorporates the predictions of the outcome:

$$\text{Predicted} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{1 - \hat{\mathbb{P}}(A_i = 1 \mid W_i^C)}\right)\hat{\mathbb{E}}(Y_i \mid W_i^C). \quad (3.10)$$

The predicted component of CARE–IPW converges to zero when the propensity scores are consistently estimated. For each level of exposure, the numerator (with the true conditional probability of exposure given confounding covariates) and the denominator (with the propensity scores) converge to one, leaving the difference in predictions $\hat{\mathbb{E}}(Y \mid W^C)$ for the two exposure levels across all units. Since the predictions do not vary by exposure level (they are independent of the exposure A), the predicted component converges to zero. Thus, CARE–IPW is consistent for the target statistical parameter Ψ, which equals the ATE when the identifiability assumptions hold.

Each of the estimators described in this paper can be characterized as special cases of CARE–IPW. CARE–IPW reduces to CARE when the propensity scores are estimated with the empirical probability of exposure. CARE–IPW reduces to IPTW when the predicted values of the outcome are all zero. CARE–IPW reduces to the unadjusted estimator when both of the above conditions are met.

As shown in Appendix B.2, under standard regularity conditions,[194] CARE–IPW is asymptotically normal and its asymptotic variance is equivalent to the sample variance divided by n. We can then construct Wald-type 95% confidence intervals using the asymptotic variance. Since CARE is a special case of CARE–IPW, its asymptotic variance and confidence intervals can be similarly estimated.

3.4 Simulation Studies

In this section, we use simulations to compare the performance of the unadjusted, IPTW, CARE, and CARE–IPW estimators of the statistical parameter Ψ (3.2). We

consider two data generating processes. In Simulation 1, we consider a simple process in which the effect of a binary exposure on a binary outcome is confounded by two covariates. Then in Simulation 2, we consider a more realistic process, designed to resemble the motivating data application. In all settings, there is no unmeasured confounding and positivity holds by design; estimates can, therefore, be interpreted causally.

We compared the performance of the estimators using bias, Monte Carlo standard error, average standard error, confidence interval coverage, power, and type I error. Each estimator made an estimate $\hat{\Psi}_s$ for the statistical parameter Ψ in simulation s, $s = 1, \ldots, S$, with a variance ν_s, and a confidence interval based on the estimate and the variance. Bias is the difference between the average estimate and the statistical parameter $\frac{1}{S}\sum_{s=1}^{S} \hat{\Psi}_s - \Psi$. Monte Carlo standard error is the standard error in the estimates across simulations $\sqrt{Var(\hat{\Psi}_{1:S})}$. Average standard error is the mean of the estimated standard errors across simulations $\frac{1}{S}\sum_{s=1}^{S} \sqrt{\nu_s}$. Confidence interval coverage is the observed proportion of 95% confidence intervals that covered the statistical paramater Ψ across all simulations. Power is the observed proportion of simulations that the estimator rejected the null hypothesis of no exposure effect when there was an exposure effect. Type I error is the observed proportion of simulations that the estimator rejected the null hypothesis of no exposure effect when the null hypothesis was true.

All simulations were run using R version 3.4.3.[154] Simulations were run in parallel on 15 cores on a remote server. To maintain reproducibility and to make sure that the same samples were drawn for each scenario (with and without an effect in randomized and observational settings) across simulations, we set a seed for the random number generator for each simulation based on the simulation number. The code used for this project can be found on GitHub.

3.4.1 Simulation 1

3.4.1.1 Setup

To observe the finite sample properties of CARE and CARE–IPW relative to those of the unadjusted and IPTW estimators, we designed a synthetic simulation with binary exposures and outcomes.

Consider an experiment with 96 units. For each unit in the sample, we generated four independent baseline covariates: $W1 \sim Normal(0,1)$, $W2 \sim Normal(0,1)$, $W3 \sim Uniform(0,1)$, and $W4 \sim Bernoulli(p = 0.5)$. We simulated a randomized trial where the exposure A was assigned with probability 0.5 as well as an observational setting where the exposure was assigned with a probability given by $logit^{-1}[1 - 0.75^{*}W1 - 2^{*}W4 + 0.5^{*}W2]$. Each unit's outcome was generated as

$$Y = \mathbb{I}\big(U_Y < logit^{-1}[-0.25 + 0.5^{*}W1 - 1^{*}W3 + 2^{*}W4 - 1.25^{*}A - 0.5^{*}A^{*}W3]\big)$$

with $U_Y \sim Uniform(0,1)$. We generated the counterfactual outcomes for each unit, $Y(1), Y(0)$, by deterministically setting the exposure to $A = 1$ and $A = 0$, respectively. The average treatment effect was calculated by taking the mean difference in the counterfactual outcomes for a population of 100,000 units. We also simulated a scenario under the null hypothesis of no exposure effect by setting the counterfactual outcome with the exposure $Y(1)$ equal to the counterfactual outcome without the exposure $Y(0)$.

We implemented the unadjusted estimator as the difference in average outcomes between exposed and unexposed units. When estimating the propensity score, required for the IPTW estimator and CARE–IPW, we used a logistic regression with main terms for $W1$ and $W4$, which are the confounders in the observational setting. For the outcome prediction, which is required for CARE and CARE–IPW, we used a logistic regression with main terms for $W1$, $W3$, and $W4$, which corresponds to the correctly

specified regression under the null. Statistical inference for all estimators was based on the estimated influence curve, as described in Appendix B.2.

3.4.1.2 Results

Table 3.1 provides a comparison of the performance of the estimators over 5,000 repetitions of the simulation. When there was an effect, the intervention A led to a 28.1% average reduction in the outcome. All estimators were unbiased in the randomized trial setting, as the confidence interval coverage for each algorithm was at or above the nominal level of 95%. By estimating the known exposure mechanism, the IPTW estimator was more precise than the unadjusted estimator, as observed by having smaller Monte Carlo standard errors. However, this precision gain did not translate into improved power for the IPTW estimator due to its conservative variance estimation. Within the estimating equation framework, an improvement in power was achieved with both CARE and CARE–IPW, which included covariates when predicting the outcome. Under the null, all estimators were unbiased and had good to conservative Type I error control in a trial setting.

When there was an effect in the observational setting, the unadjusted estimator was biased with low confidence interval coverage. By controlling for confounders when predicting the outcome, CARE had less bias and greater confidence interval coverage, though still less than the nominal level of 95%. However, through consistent estimation of the propensity score and thereby control for the confounders, both the IPTW estimator and CARE–IPW were unbiased and had nominal to conservative confidence interval coverage. CARE–IPW achieved greater statistical power than the IPTW estimator by having less Monte Carlo and average standard error.

When there was no effect in an observational setting, the unadjusted estimator was again biased with low confidence interval coverage. In contrast, the IPTW estimator, CARE, and CARE–IPW were unbiased with nominal to conservative Type I error

Trial	Exposure	Estimator	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
RCT	Effect	CARE–IPW	0.003	0.092	0.092	94.5%	85.4%
		CARE	0.008	0.090	0.090	94.4%	85.3%
		IPTW	-0.001	0.094	0.148	99.7%	46.2%
		Unadj	-0.002	0.101	0.101	94.3%	78.1%
RCT	Null	CARE–IPW	0.000	0.093	0.090	94.1%	5.9%
		CARE	0.000	0.091	0.089	94.4%	5.6%
		IPTW	0.000	0.095	0.167	99.9%	0.1%
		Unadj	-0.000	0.104	0.103	94.4%	5.6%
Obs	Effect	CARE–IPW	0.000	0.115	0.115	94.5%	71%
		CARE	0.062	0.082	0.081	87.4%	75.6%
		IPTW	-0.005	0.126	0.164	98.7%	44.6%
		Unadj	-0.197	0.088	0.089	41.7%	100%
Obs	Null	CARE–IPW	-0.004	0.107	0.102	94.1%	5.9%
		CARE	-0.003	0.079	0.087	96.7%	3.3%
		IPTW	-0.005	0.124	0.197	99.6%	0.4%
		Unadj	-0.219	0.100	0.099	39.7%	60.3%

Table 3.1. Results for the effect estimators in Simulation 1 by trial type and exposure. The covariate-adjusted residuals estimator (CARE) used a logistic regression with $W1$, $W3$, and $W4$ to predict the outcome. The inverse-probability of treatment weighting (IPTW) estimator uses a logistic regression with $W1$ and $W4$ to estimate the propensity scores. CARE with inverse probability weighting (CARE–IPW) the same regression as CARE to predict the outcome and the same regression as IPTW to estimate the propensity scores.

control. We note that under the null, CARE is expected to be consistent if the outcome is correctly predicted, which it was here.

Altogether these simulations confirm the theoretical properties, described in Section 3.3.

3.4.2 Simulation 2

3.4.2.1 Setup

Here we observe the performance of CARE and CARE–IPW in a simulation inspired by the bednet cluster-randomized trial.[79] The baseline covariates and the outcome were generated using distributions based on the case study data. As in Simulation 1,

we ran 5,000 repetitions of four different scenarios: two levels of exposure effect, with an effect and under the null of no exposure effect, for exposures that were randomly assigned and for exposures that were assigned as a function of covariates.

We generated data for 96 clusters for each simulation. The number of person-years (M) for each cluster were drawn from a Poisson distribution with the same mean as observed in the case study. Baseline covariates for the average age in months ($W1$) and the percent of children who were female ($W2$) were drawn from normal distributions with the same mean and variance as observed in the case study. An additional covariate, prior childhood mortality rate ($W3$), was generated using draws from a negative binomial distribution with a similar mean (μ) and overdispersion parameter (r) as seen in the unexposed group of the case study. The counterfactual for childhood mortality without bednets ($Y(0)$) was drawn from a negative binomial distribution based on the number of person-years M and the baseline covariates $W1$, $W2$, and $W3$ for each cluster. For trials with an exposure effect, the counterfactual for childhood mortality with bednets ($Y(1)$) was a deterministic formula based on $Y(0)$ and the prior childhood mortality rate $W3$, which translates into a reduction of about 12 deaths per thousand person-years. For trials without an exposure effect, the two counterfactuals were equivalent $Y(1) = Y(0)$. The data generating equations were as follows:

$$
\begin{aligned}
M(\text{Person-years}) &\sim \text{Poisson}(347) \\
W1(\text{Average age in months}) &\sim \text{Normal}(25.3, 1) \\
W2(\text{Percent female}) &\sim \text{Normal}(0.5, 0.03) \\
W3(\text{Prior mortality rate}) &\sim \text{NegBinom}(0.04^{*}M, r = 16)/M \\
\mu &= (0.378 - 0.01^{*}W1 - 0.25^{*}W2 + W3)^{*}M \\
Y(0) &\sim \text{NegBinom}(\mu, r = 16) \\
Y(1) &= Y(0), \text{ under the null} \\
&= Y(0) - (0.01 + .1^{*}W3)^{*}M, \text{ o.w.}
\end{aligned}
$$

In the randomized setting, exactly 48 clusters were assigned to both exposure levels in each simulation. In the observational setting, the probability of exposure depended on the prior childhood mortality rate:

$$\mathbb{P}(A = 1 \mid W3) = \text{Binomial}(logit^{-1}(60^{*}(W3 - .04))).$$

Thus prior childhood mortality rate $W3$ was a common cause of both the exposure and the outcome in the observational simulations. The observed childhood mortality Y was the realization of the counterfactual for the observed exposure level.

In our data generating process, there were no unmeasured confounders and the positivity assumption held by design. Therefore, the statistical parameter Ψ (3.2) identified the ATE. To calculate the ATE, we generated a population of 100,000 clusters with the same process that generated the data for each simulation, then took the average difference between the counterfactual outcomes.

For each simulation, we made estimates of the statistical parameter Ψ with the unadjusted estimator, the IPTW estimator, CARE, and CARE–IPW using several different approaches to predict the outcomes and estimate the propensity scores. When predicting the outcome Y (with an offset for person-years M) for CARE and

CARE–IPW, we used a main terms negative binomial generalized linear regression with all baseline covariates ($W1$, $W2$ and $W3$), as well as a pair of main terms Poisson generalized linear regressions: one using only prior childhood mortality ($W3$, the only confounder in the observational setting) and the other using only age and sex ($W1$ and $W2$, as in the real trial). For estimating the propensity scores for the IPTW estimator and CARE–IPW, we used main terms logistic generalized linear regressions with three combinations of baseline covariates: all baseline covariates ($W1$, $W2$, and $W3$), only prior childhood mortality ($W3$), and only age and sex ($W1$ and $W2$).

3.4.2.2 Results

Across the large generated population with an exposure effect, the ATE was -11.7 deaths per thousand person-years (from $Y(1)$ =28.5 to $Y(0)$ =40.3). In the simulations of observational settings, there were an average of 47.8 (range: 27 to 64) clusters assigned to the exposed group.

In the randomized trial simulations (Table 3.2), all of the estimators had low bias and high confidence interval coverage. CARE, CARE–IPW, and the IPTW estimator had less Monte Carlo standard error than the unadjusted estimator. When there was an exposure effect, the CARE and CARE–IPW estimators had more statistical power than the unadjusted and IPTW estimators. The average standard errors for the CARE and CARE–IPW estimators were smaller than those of the unadjusted and IPTW estimators. This simulation suggests that the CARE and CARE–IPW estimators offer improvements to the unadjusted and IPTW estimators in randomized trials due to their similar levels of bias and confidence interval coverage, greater statistical power, and smaller variance. Results for simulations that were limited to using age $W1$ and sex $W2$, as in the case study, are presented in the Supplementary Materials B.3.5.

In the observational simulations, CARE–IPW made unbiased forecasts with low variability despite strong confounding between the exposure and the outcome, com-

Exposure	Estimator	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
Effect	CARE–IPW	0.00	4.66	4.55	94.1%	72.6%
	CARE	0.13	4.59	4.51	94.2%	72.4%
	IPTW	0.00	4.64	8.88	100%	10.9%
	Unadj	0.01	5.35	5.24	94.4%	60.6%
Null	CARE–IPW	-0.02	4.73	4.62	94.3%	5.7%
	CARE	-0.02	4.67	4.59	94.5%	5.5%
	IPTW	0.00	4.72	9.98	100%	0%
	Unadj	0.01	5.53	5.42	94.4%	5.6%

Table 3.2. Simulation results for the effect estimators in randomized trials with and without an exposure. The covariate-adjusted residuals estimator (CARE) uses a Poisson regression with the prior childhood mortality $W3$ as a covariate to predict the outcome. The inverse-probability of treatment weighting (IPTW) estimator uses a logistic regression with $W3$ as a covariate to estimate the propensity scores. CARE–IPW the same regression as CARE to predict the outcome and the same regression as IPTW to estimate the propensity scores.

paring favorably to other estimators. Both with an effect and under the null, the unadjusted estimator had large bias and poor confidence interval coverage, indicating that there was confounding between the exposure and the outcome. When the estimators only accounted for the confounding covariate, prior childhood mortality $W3$ (Table 3.3), CARE–IPW had the least bias and the most statistical power when there was an effect; CARE had the least Monte Carlo and average standard error, but the most bias when there was an effect; the IPTW estimator had low bias, but most Monte Carlo and average standard error; and all three estimators had high confidence interval coverage. More results from the observational simulations are in the Supplementary Materials B.3.5.2.

3.5 Case study: bednets in Ghana cluster-randomized trial

In this section, we first reproduced the findings of Hayes and Moulton, who compared CARE to the unadjusted estimator for a cluster-randomized trial in northern

Exposure	Estimator	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
Effect	CARE–IPW	-0.02	5.08	4.98	94.9%	67.3%
	CARE	1.45	4.36	4.56	94.9%	61.9%
	IPTW	0.26	6.10	10.08	99.9%	5.1%
	Unadj	8.59	5.27	5.20	61.6%	10%
Null	CARE–IPW	-0.11	5.28	5.03	94.6%	5.4%
	CARE	0.13	4.40	4.62	95.8%	4.2%
	IPTW	0.27	6.23	11.05	100%	0%
	Unadj	9.30	5.44	5.37	58.3%	41.7%

Table 3.3. Simulation results for the effect estimators in observational studies with and without an exposure effect. The covariate-adjusted residuals estimator (CARE) uses a Poisson regression with the confounding covariate prior childhood mortality $W3$ to predict the outcome. The inverse-probability of treatment weighting (IPTW) estimator uses a logistic regression with the confounder $W3$ as a covariate to estimate the propensity scores. CARE–IPW the same regression as CARE to predict the outcome and the same regression as IPTW to estimate the propensity scores.

Ghana that measured the impact of impregnated bednets on child mortality.[10, 79, 16] Then we applied the IPTW estimator and CARE–IPW on the same data and discussed the results.

3.5.1 Setup

In the original analysis, the researchers estimated the unadjusted and covariate-adjusted mortality rates for the exposed and unexposed groups and compared them using the t-test. For the unadjusted estimator, the observed mortality rate (*i.e.* the number of deaths per thousand followup-years) was calculated for each cluster. The unadjusted estimate of the exposure effect was equal to the difference in the average observed mortality rate between the exposure levels.

In the covariate-adjusted analysis, the researchers used a Poisson generalized linear regression for mortality rate on the individual-level data using age and sex as covariates while disregarding the cluster assignment and the exposure level. From this regression, they predicted the expected mortality rate per follow-up year for each child,

which they then aggregated into cluster-level predicted mortality rates per thousand followup-years. The researchers found the residuals by taking the difference between the predicted and observed mortality rates for each cluster. The CARE estimate of the exposure effect was equal to the difference in the average of the residuals between exposure levels. Hayes and Moulton used a t-test to generate confidence intervals and conduct hypothesis testing.

We reproduced the analysis above and extended it to include IPTW and CARE–IPW. While our estimates of the average treatment effect for CARE and the unadjusted estimator were the same as in Hayes and Moulton, we estimated the variance using the influence curve-based methods described in Appendix B.2 which yielded slightly different confidence intervals and p-values. The IPTW and CARE–IPW estimators required propensity scores. We estimated these using a main terms logistic generalized linear regression for the exposure at the cluster level using average age in months and percent of children who were female as covariates. These covariates may be relevant as younger children are more vulnerable than older children and young males typically have a higher mortality rate than young females. For CARE–IPW, we used the same predicted values of the outcome from the individual-level regression as used for CARE. We estimated the average treatment effect and variance for CARE and the IPTW estimator using the methods outlined in Section 3.3.3.

3.5.2 Results

The IPTW and CARE–IPW estimates of the exposure effect were larger than the estimates from the unadjusted estimator or CARE (Figure 3.3). As in the original analysis, we found a mortality rate difference between the exposed group and the unexposed group of -3.95 (95% CI: -8.46, 0.56; p-value: 0.09) per thousand followup-years using the unadjusted estimator and -4.26 (-8.67, 0.15; p-value: 0.06) per thousand followup-years using CARE. Using IPTW, the estimated mortality rate difference

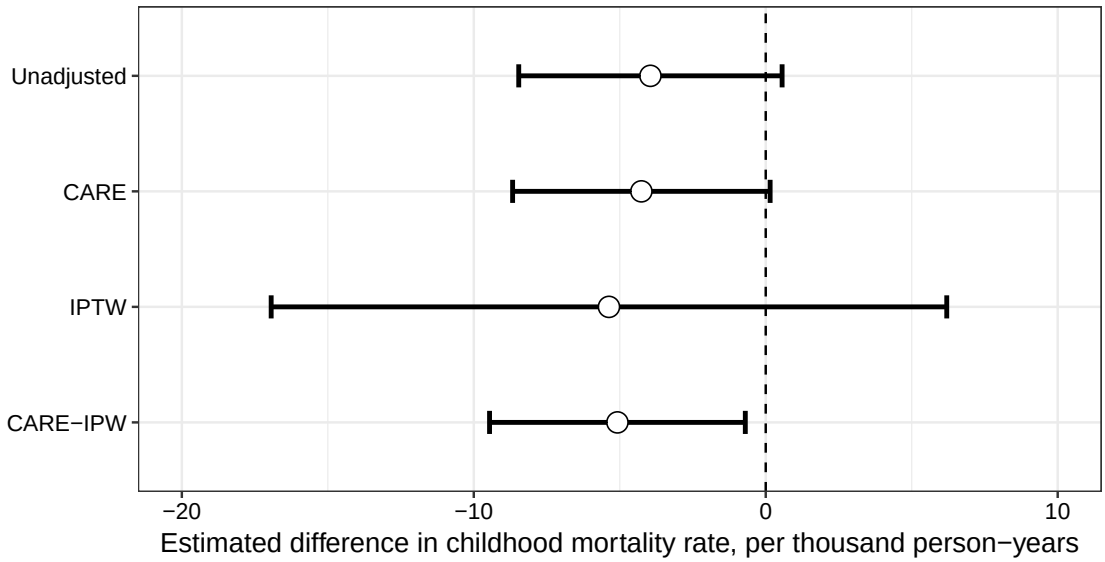

Figure 3.3. The estimates and 95% confidence intervals for the effect of allocating bednets on childhood mortality rate per thousand person-years. The four estimators used are the unadjusted estimator, the covariate-adjusted residuals estimator (CARE), the inverse probability of treatment weighting estimator (IPTW), and CARE with inverse probability weighting (CARE–IPW). Whie all estimates indicate that bednets caused a reduction in childhood mortality, the CARE and CARE–IPW estimates were more precise than those of the unadjusted and IPTW estimators.

was -5.37 (-16.94, 6.2; p-value: 0.36) per thousand followup-years. For CARE–IPW the mortality rate difference was -5.08 (CI: -9.46, -0.7; p-value: 0.02) per thousand followup-years. As in the simulation study, the standard error for the CARE and CARE–IPW estimates were less than those of the unadjusted and IPTW estimates. While IPTW had the largest estimated effect size, it also had the largest variance of any estimator. The estimate made by CARE–IPW was larger than either the unadjusted estimator or CARE and had the smallest variance of any estimator.

3.6 Discussion

In this paper, we provided a non-parametric statistical justification for the covariate-adjusted residuals estimator (CARE) in randomized and observational settings, pro-

posed a novel estimator, the covariate-adjusted residuals estimator with inverse probability weighting (CARE–IPW), and supported the theory with a simulation study and an application to a cluster-randomized trial. Specifically, we proved that CARE is consistent for the average treatment effect (ATE) in randomized studies, where there is no unmeasured confounding or violations of the positivity assumption by design. We also proved that CARE is not consistent for the ATE in observational settings. We developed a new estimator, CARE–IPW, which is consistent for the ATE in observational settings when the propensity scores are consistent for the true conditional probability of exposure given confounding covariates, and when there is no unmeasured confounding or violations of the positivity assumption.

The unadjusted estimator, the inverse probability of treatment weighting (IPTW) estimator, and CARE are special cases of CARE–IPW. In a randomized trial, we would expect CARE to be more efficient than the unadjusted estimator when the predictions for the outcome account for covariates that are predictive of the outcome and are imbalanced between exposure levels.

The simulation studies supported our theoretical findings and suggested some advantages to using CARE–IPW rather than CARE or the IPTW estimator. In randomized trials, CARE and CARE–IPW had similar levels of bias and confidence interval coverage to the comparison estimators, but with greater levels of statistical power. In observational settings, CARE–IPW was consistent for Ψ when accounting for the confounding covariate in the propensity score model and had greater statistical power and less variability than the IPTW estimator. CARE had more bias than CARE–IPW or the IPTW estimator in observational settings with an intervention effect.

While CARE–IPW improves on CARE and IPTW, it is not a "double robust estimator", such as targeted maximum likelihood estimation[195] and augmented inverse probability weighting.[173] A double robust estimator is consistent for Ψ if

either the outcome predictions (which often include the exposure as well as baseline covariates) or the propensity scores are consistently estimated and is the most efficient estimator if both are. Similar to the IPTW estimator, CARE–IPW is consistent for Ψ if and only if the propensity scores is consistently estimated. CARE–IPW may improve efficiency over the IPTW estimator by making predictions of the outcome.

One advantage to using CARE–IPW rather than another method is that researchers do not need to specify the relationship between the exposure and the outcome. This can be beneficial when there is a complex relationship between the exposure and outcome, such as multiple non-linear interactions with other covariates that augment the strength of the exposure.

As with the IPTW estimator, CARE–IPW may have stability issues when estimated propensity scores approach zero or one.[150] This could be resolved in one of a couple ways. Stabilized weights could be used to scale propensity scores away from zero and one.[164] Alternatively, propensity scores could be replaced by incremental propensity scores which relax the positivity assumption by looking at the effect of an intervention when propensity scores are uniformly increased and decreased across all observations.[104]

3.7 Acknowledgements

We thank Mark van der Laan for his expert advice.

This project was funded by NIH NIAID grant 1R01AI102939 and NIGMS grant R35GM119582. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the views of the National Institutes of Health or the National Institute of General Medical Sciences. The funders had no role in study design, data collection and analysis, decision to present, or preparation of the presentation.

CHAPTER 4

INCORPORATING FORECASTS INTO ESTIMATES OF THE AVERAGE TREATMENT EFFECT WITH AN APPLICATION TO ZIKA EMERGENCY OPERATIONS CENTERS IN THAILAND

4.1 Introduction

In this paper, we investigate whether forecasts can improve estimates of the effect of an exposure in observational settings. Observational studies can provide useful evidence about the effect of an exposure or intervention,[18, 71, 146, 188] but researchers need to account for covariates that can bias effect estimates.[166] Identifying these covariates can be difficult when there are a large number of potential causes and a small number of observations.[8, 121, 201] Using a forecasting paradigm to perform covariate selection may improve effect estimation in specific situations.

Consider a scenario where we observe outcome $Y_{i,t}$ for $i = 1, \ldots, n$ units at each time $t = 1, \ldots, T$. For each unit-time, we observe a p-length set of baseline covariates $\mathbf{X}^p_{i,t}$ that occur prior to the outcome. For example, each time t could be a year with the baseline covariates occurring early in the year and the outcome occurring later in the year; both are observed at different moments within the same time period. At exposure time T, a subset of the units receive an exposure that occurs after the baseline covariates and prior to the outcome; A_i is a binary variable that indicates whether unit i received the exposure at time T. A k-length subset of the baseline covariates are the common cause covariates $\mathbf{X}^k_{i,t}$, which causally influence both the exposure (at time T) and the outcome (across all times t). For short, we denote the baseline covariates at the exposure time $\mathbf{X}_i = \mathbf{X}^p_{i,T}$.

The causal parameter that we want to estimate in this scenario is the average treatment effect (ATE). There are two potential outcomes $Y_i(A = a)$ for each unit i at the exposure time T, corresponding to the counterfactuals where the unit is exposed $Y(1)$ and unexposed $Y(0)$. The ATE is the average difference between the potential outcome when a unit is exposed and the potential outcome when a unit is unexposed across all of the units in our population:

$$\text{ATE} = \mathbb{E}(Y(1) - Y(0)). \tag{4.1}$$

Under a set of assumptions, the ATE can be identified using the expectation of the difference in conditional means with and without the exposure, which we denote as the statistical parameter $\Psi = \mathbb{E}_{\mathbf{X}}\big[\mathbb{E}(Y \mid A = 1, \mathbf{X}) - \mathbb{E}(Y \mid A = 0, \mathbf{X})\big]$.[167, 164, 150] The first assumption is that the potential outcomes for each unit are dependent only on the baseline covariates $\mathbf{X}_i$ and the exposure A_i for that unit at the exposure time T, without interference from exposures or outcomes from other units. The second assumption is that there is no unmeasured confounding, so that the potential outcomes and the exposure are independent given our baseline covariates $Y(a) \perp\!\!\!\perp A \mid \mathbf{X}$. The third assumption is that of positivity, whereby each strata of covariates have a non-zero probability of assignment to both exposure groups (for a binary exposure) $0 < \mathbb{P}(A = 1|\mathbf{X}) < 1$.

The positivity assumption puts a limit on the number of observations that can be used to estimate the statistical parameter Ψ. All of the observations that occur prior to the exposure time T have no probability of being exposed and must be excluded from the analysis. Thus, the analysis is restricted to using the observations at the exposure time T, which has a sample size of n. If past observations confound the relationship between the potential outcome $Y(a)$ and the exposure A, then those need to be included as covariates. This would increase the number of baseline covariates $\mathbf{X}_i$ and make identifying the true subset of common cause covariates $\mathbf{X}^k_{i,T}$ more difficult.

We propose using a forecasting paradigm to perform covariate selection using all of the observations prior to exposure time T. Forecasting models are not restricted to using observations that have a non-zero probability of exposure. Instead, all of the observations prior to time T (a total of $(T-1)^*n$ observations) can be used in a covariate selection procedure to train a forecasting model for predicting the outcome $Y_{i,t}$ with the baseline covariates $\mathbf{X}^p_{i,t}$. This forecasting model would then make predictions $\tilde{Y}_i$ of the outcome at exposure time T. Subsequently, these predicted values $\tilde{Y}_i$ would be used as a covariate for an effect estimator to estimate the statistical parameter Ψ. With this paradigm, we can use a larger sample of observations to estimate the relationship between the baseline covariates $\mathbf{X}^p_{i,t}$ and the outcome $Y_{i,t}$, while also using a parsimonious model to make the effect estimates.

In this paper, we use two simulations to examine whether using forecasts in place of traditional methods improves effect estimation. To do this, we make effect estimates with two estimators and compare the estimates that use forecasting to those using two approaches that do not use the forecasted values. We then implement the most suitable estimators to an application in an infectious disease setting.

4.2 Case study

As a running example, we use an intervention implemented by the Thailand Ministry of Public Health (MOPH) in response to the Zika virus global pandemic of 2016.

Dengue and Zika are flaviviruses that actively circulate in Thailand and are spread by *Aedes* mosquitoes.[161, 73] Dengue, in particular, has been a major public health priority for the MOPH, which started tracking the disease by province in 1968. A dengue infection can cause dengue hemorrhagic fever (DHF), a severe disease that can lead to organ failure or even death.[161] Since 2000, there has been an average of 41,795 reported DHF cases each year in Thailand, though wide-spread epidemics can

cause over 100,000 DHF cases and require rapid, large-scale government response.[39] By contrast, Zika virus has historically been associated with a less serious immune response, similar to the mild manifestation of dengue, and only with the 2016 pandemic became a public health priority.[136] There were reported Zika cases in Thailand prior to 2016, however the virus was infrequently tested for due to its relative mildness and similarity to dengue; from the start of surveillance by the MOPH in 2012 to 2015, there was an average of 5 Zika cases reported per year.[169, 203, 143]

The Zika virus global pandemic changed the MOPH protocol for handling Zika cases. During the pandemic, Zika spread to over 80 countries and territories and was associated with a rise in neurological complications, specifically severe microcephaly for newborns and Guillain–Barré syndrome in adults.[3, 106, 94, 41, 31, 101] This led the World Health Organization to announce a public health emergency of international concern from February to November of 2016.[73] That year, the MOPH set up emergency operations centers (EOCs) in districts with reported cases to test infected people, monitor high-risk populations, and recommend interventions to local officials.[143, 2] At the end of June, there were 97 reported Zika cases in 10 provinces and by the end of the year there were 1,114 reported Zika cases in 42 provinces (Figure 4.1).[1, 205]

While the EOCs were deployed to stem the spread of Zika, they might have affected DHF incidence as well. Determining whether the EOCs reduced Zika incidence is difficult because Zika incidence was inconsistently reported prior to 2016 and the EOCs were only deployed to regions with reported Zika incidence. Furthermore, part of the EOC's mission was to seek out other Zika cases, so the deployments might appear to have increased Zika incidence rather than to have decreased it. Using DHF incidence in place of Zika cases could mitigate these issues. With the MOPH's history of dengue surveillance and the fact that DHF often requires hospitalization for survival, we expect that there would be less variation in DHF reporting rates over time and

that EOCs would not have discovered many new DHF cases that would not have been reported anyways. Since Zika and dengue share a vector, *Aedes* mosquitoes, we assume that the viruses arise under similar conditions and that interventions that attempt to inhibit their transmission would have similar effects on their incidence. Thus, we have a natural experiment where we can measure whether the Zika EOCs were associated with subsequent reductions in DHF incidence.

This case study is a good example of a situation where effect estimation is difficult and forecasts may improve inference for this effect estimate. To make accurate estimates of effect size, we would prefer to have a large sample size, balanced numbers of exposed and unexposed units, and a few well-known causes of both the intervention and the outcome. In this situation, we only have 76 observations (1 outcome per province) at our exposure time, with many more unexposed than exposed provinces (66 provinces with no EOCs vs. 10 provinces with EOCs). To further complicate matters, there are many potential common causes of both the intervention and the outcome, whose true relationships may be concealed by dengue's complex transmission dynamics. Many dengue infections are asymptomatic and multi-year cross-protection between strains can conceal the true population susceptibility. If, for instance, provinces with high susceptibility to dengue in a given year coincidentally had more rainfall, a model based only on that year may disproportionately associate rainfall with DHF incidence. By looking across multiple years, we may be able to identify the true relationships between potential common causes and DHF incidence. In Chapter 2, we made forecasts for DHF incidence for Thai provinces.We adapt those forecasts for use in these effect estimation exercises.

4.2.1 Data

The exposure for this natural experiment was the deployment of an emergency operations center (EOC) to a Thai province. The Bureau of Emerging Infectious

Diseases released a guidebook with protocols for EOCs after a confirmed Zika case.[2] One objective of the EOCs was to suggest prevention and control strategies to local, provincial, and regional authorities; however we do not know the extent to which these strategies were implemented. Since official MOPH documents both state that EOCs should be deployed after reported Zika cases and that 10 provinces reported Zika cases prior to 29 June 2016, we assumed that EOCs were deployed to these provinces. Thus, we used EOC deployment as our exposure on an intention to treat basis.

The outcome of interest for this case study was the provincial DHF incidence rate, per 100,000 population. We obtained DHF data from the MOPH and provincial population data from the National Statistics Office of Thailand. Since the exposure occurred prior to 29 June 2016, we aggregated the DHF incidence from July through December of each year. We used data from 2000 to 2015 to train our forecasts for 2016. There was considerable temporal and spatial variation in DHF incidence rates over that time period (Figure 4.2).

There were several potential sources of common causes between the Zika EOCs and provincial DHF incidence rates. For infectious disease transmission to take place, there needs to be an environment for transmission, infected people, and susceptible people. To represent the environment for transmission, we used monthly temperature and rainfall data on 0.5x0.5 latitude-longitude resolution from the Earth System Research Laboratory at the National Oceanic and Atmospheric Administration.[50, 5] Since the Zika EOCs were deployed between February and June, we restricted the baseline covariates to time frames prior to February 2016. Thus, the baseline weather covariates were monthly temperature and rainfall for the November, December, and January immediately preceding the exposure. While the pre-intervention DHF incidence rate (aggregated monthly from November through January) might not have been a direct cause of a Zika EOC deployment, it could indicate that there was a suitable environment for disease transmission by *Aedes* mosquitoes. Similarly, population

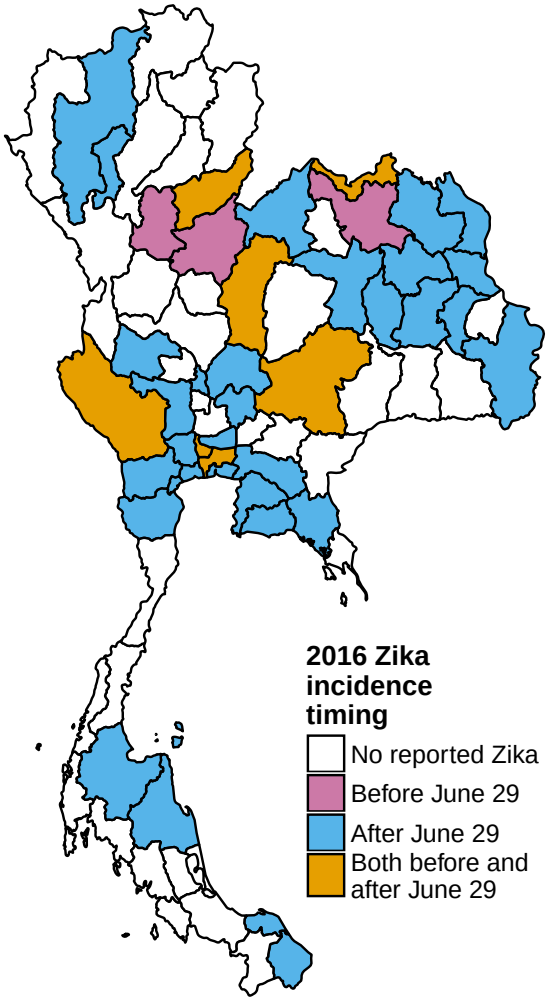

Figure 4.1. The location and timing of Zika incidence in Thailand over the course of 2016.

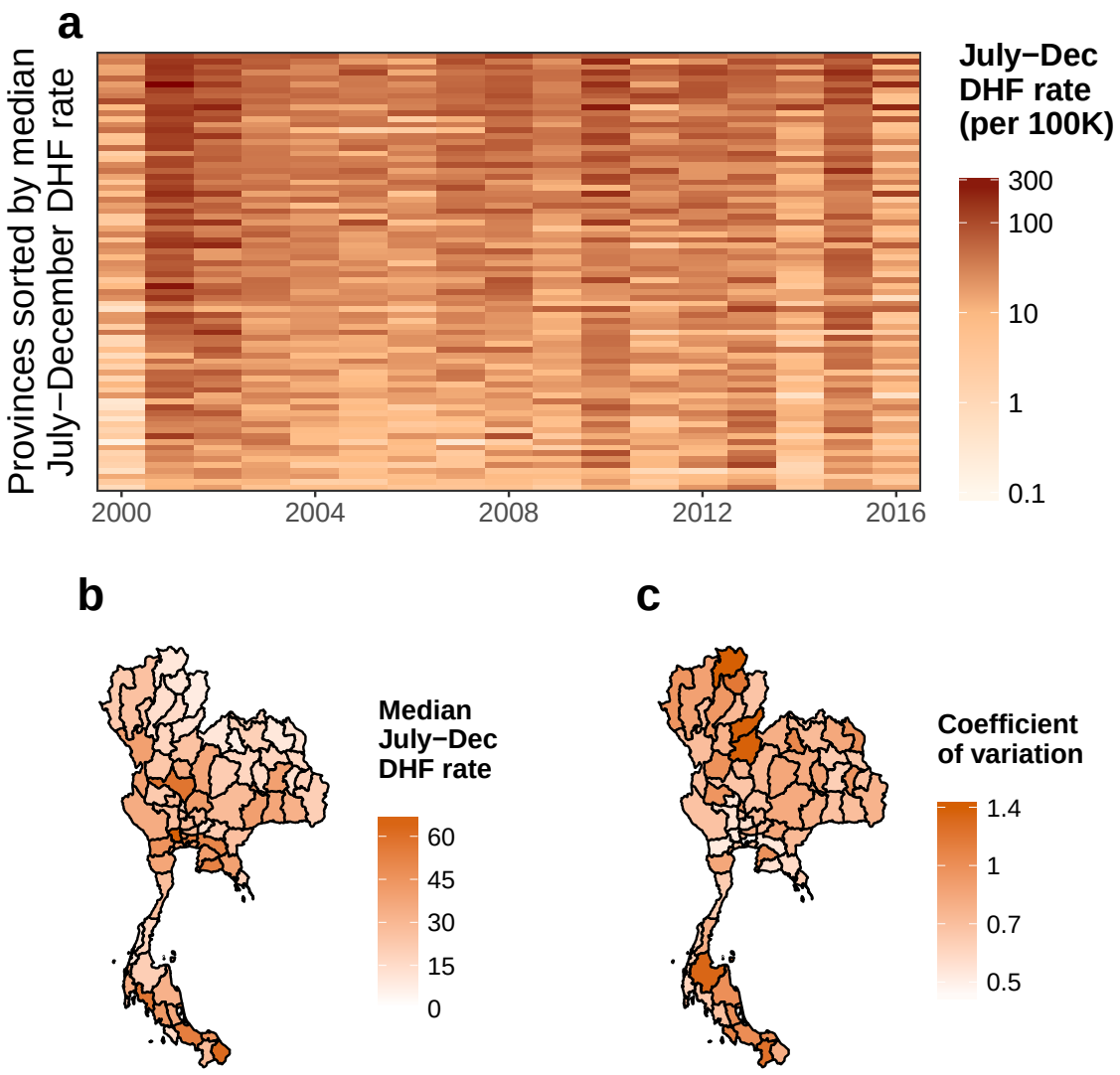

Figure 4.2. The temporal and spatial distribution of annual dengue hemorrhagic fever (DHF) incidence rates in Thailand. **(a)** The annual DHF incidence rate, per 100,000 population, for each Thai province and year used in this study. **(b)** The median annual DHF incidence rate, per 100,000 population, for each province from 2000-2014. **(c)** The coefficient of variation (standard deviation divided by the mean) of the annual DHF incidence rate for each province.

susceptibility to dengue is dependent on past infections, which were not direct causes of Zika incidence. However, historical DHF incidence rates could indicate which provinces have the environmental capacity for dengue outbreaks, and thus Zika infections. We represented these historical DHF incidence rates with the July through December incidence rates for each year from 2000 to 2015, as well as the minimum, median, and maximum incidence rates for each province over that span. The number of people in a province was a weak predictor of the DHF incidence rate, since the rate is scaled by population. However, provincial population was strongly predictive of the deployment of Zika EOCs because a larger population offered more opportunities for reported Zika cases than a smaller population did. All of these relationships are depicted in our causal diagram (Figure 4.3).

4.3 Methods

4.3.1 Forecasting paradigm

For both of the simulation studies and the case study, we selected the forecasting model using a procedure similar to the one used in Chapter 2. In that chapter, the data was separated into a training phase, from 2000 to 2009, and a testing phase, from 2010 to 2014. In the training phase, a forward-backward covariate selection process with leave-one-year-out cross validation was used to select two models to make forecasts of DHF incidence. The first selected model had the least cross-validated mean absolute error in the training phase. However, after cross validating over 200 different forecasting models, the model with the least cross-validation error was likely to have overfit to the training phase data. Thus, a parsimonious model, the model with the least covariates with a cross-validated mean absolute error within one standard error of the first model, was also selected. We then used these two models to prospectively forecast the testing phase, first using all of the data from 2000-2009 to forecast 2010 then using all of the data from 2000-2010 to forecast 2011, and so on. In that chapter,

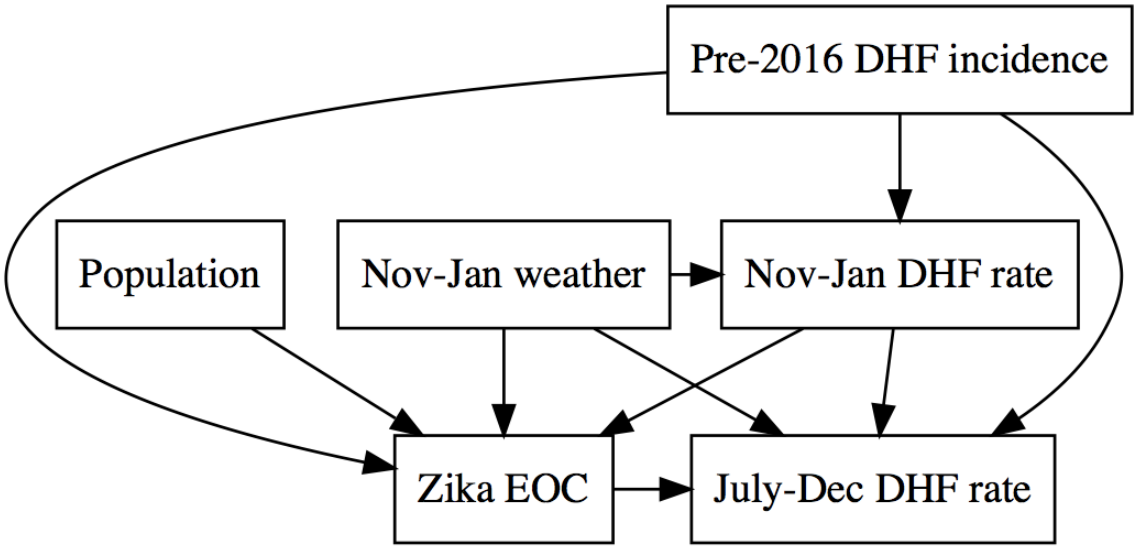

Figure 4.3. A diagram depicting the causal model for estimating the effect of deploying Zika emergency operations centers (EOCs) on July-December dengue hemorrhagic fever (DHF) incidence rate. The model includes covariates that account for provincial population, weather, and prior DHF incidence.

the parsimonious model had less mean absolute error, higher prediction interval coverage, and better outbreak detection in the testing phase than the model with the least training phase error. This model selection procedure was adapted for use with each simulation and the case study.

4.3.2 Estimators

In our simulations and case study, we estimated the statistical parameter Ψ with several effect estimators.

The unadjusted estimator is the difference between the mean values of the exposed and unexposed units:

$$\hat{\Psi}^{unadj} = \frac{1}{\sum_{i=1}^{n} \mathbb{I}(A_i = 1)} \sum_{i=1}^{n} \mathbb{I}(A_i = 1) Y_{i,T} - \frac{1}{\sum_{i=1}^{n} \mathbb{I}(A_i = 0)} \sum_{i=1}^{n} \mathbb{I}(A_i = 0) Y_{i,T}. \tag{4.2}$$

This quantity is consistent for Ψ in randomized trials, but biased when there are common causes that influence both the exposure A and the outcome Y. The unadjusted estimator was used to demonstrate the bias and variability of the data in the simulations.

The inverse probability of treatment weighting (IPTW) estimator uses 'propensity scores' $\hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)$, estimates of the probability of allocation to the exposure group given baseline covariates, to estimate the statistical parameter Ψ:[167, 164]

$$\hat{\Psi}^{IPTW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)} - \frac{\mathbb{I}(A_i = 0)}{1 - \hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)} \right) Y_{i,T}. \tag{4.3}$$

The IPTW estimate $\hat{\Psi}^{IPTW}$ is consistent for the statistical parameter Ψ if the propensity scores $\hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)$ consistently estimate the true conditional probability of allocation $\mathbb{P}(A_i = 1 \mid \mathbf{X}_i)$. The IPTW estimator is especially sensitive to the positivity assumption. If any propensity score is equal to zero or one, then the estimate is

undefined; extreme propensity scores can cause unstable estimates of the statistical parameter Ψ. For this reason, we bounded the propensity scores to lie between 0.01 and 0.99.

Three methods were used to estimate the propensity scores: (1) a logistic regression model with main terms for the baseline covariates observed at the exposure time $\hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)$; (2) a lasso penalized logistic regression model using the covariates at the exposure time $\hat{\mathbb{P}}(f_{\ell_1}(A_i = 1, \mathbf{X}_i))$; and (3) a logistic regression model based on forecasted values of the outcome $\hat{\mathbb{P}}(A_i = 1 \mid \tilde{Y}_i)$.

As described in Chapter 3, the covariate-adjusted residuals estimator with inverse-probability weighting (CARE–IPW) uses propensity scores as well as conditional expectations of the outcome that are based on baseline covariates but not the exposure $\hat{\mathbb{E}}(Y_{i,T} \mid \mathbf{X}_i)$.The average contrast between the observed and predicted values (*i.e.* residuals) for each exposure group gives an estimate for Ψ:

$$\hat{\Psi}^{CARE-IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)} - \frac{\mathbb{I}(A_i = 0)}{1 - \hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)}\right)(Y_{i,T} - \hat{\mathbb{E}}(Y_{i,T} \mid \mathbf{X}_i)), \tag{4.4}$$

where the propensity scores $\hat{\mathbb{P}}(A_i = 1 \mid \mathbf{X}_i)$ are the same as those used for the IPTW estimator. As with the estimates from the IPTW estimator, the CARE–IPW estimate $\hat{\Psi}^{CARE-IPW}$ is consistent for Ψ if the propensity scores consistently estimate the true allocation probability. In some cases, the predicted values of the outcome $\hat{\mathbb{E}}(Y_{i,T} \mid \mathbf{X}_i)$ can lead to more efficient estimates than those made by the IPTW estimator.

Three methods were used to estimate the outcome for CARE–IPW: (1) a Poisson generalized linear regression model with main terms for the baseline covariates observed at the exposure time $\hat{\mathbb{E}}(Y_{i,T} \mid \mathbf{X}_i)$; (2) a lasso penalized Poisson generalized linear regression model using the baseline covariates at the exposure time $\hat{\mathbb{E}}(f_{\ell_1}(Y_{i,T}, \mathbf{X}_i))$; and (3) a Poisson generalized linear regression model based on the forecasted values

Estimator	Outcome estimation method	Propensity score method	$\mathbf{X}_i$	$\tilde{Y}_i$
Unadjusted				
IPTW(G)		GLM	✓	
IPTW(L)		Lasso	✓	
IPTW(F)		Forecasting		✓
CARE(G,G)	GLM	GLM	✓	
CARE(L,L)	Lasso	Lasso	✓	
CARE(F,L)	Forecasting	Lasso	✓	✓
CARE(L,F)	Lasso	Forecasting	✓	✓
CARE(F,F)	Forecasting	Forecasting		✓

Table 4.1. The estimators used in this paper by their fitting methods and covariates used. $\mathbf{X}_i$ are the baseline covariates for all provinces observed at the exposure time T. $\tilde{Y}_i$ are the forecasted values for all provinces at the exposure time T; with the forecasting model trained and tested on all data prior to the exposure time T.

of the outcome $\hat{\mathbb{E}}(Y_{i,T} \mid \tilde{Y}_i)$. The list of estimators used in this paper and their abbreviations are shown in Table 4.1.

Since the main terms and lasso models are restricted to be generalized linear models, we also restricted the forecasting models to be generalized linear models. Thus, we can measure the difference in using forecasted values attributable to having access to more observations and from dimension reduction, rather than to using more advanced estimation methods.

Variance estimation

We estimate the variance ν using the influence curve for each estimator. The estimators used in this paper are asymptotically linear, meaning that they are asymptotically equivalent to a sample mean of a function of the data:

$$\hat{\Psi} - \Psi = \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_{i,T}) + o_p(1/\sqrt{n}),$$

where $\varphi(Y)$ is the influence curve, also known as the influence function, and $o_p(1/\sqrt{n})$ is a random variable that converges to zero in probability.[139, 165, 103] The influence curve has mean zero and finite variance ($\mathbb{E}(\varphi(Y)) = 0$, $Var(\varphi(Y)) < \infty$) and thus, by the central limit theorem and Slutsky's theorem:

$$\sqrt{n}(\hat{\Psi} - \Psi) \xrightarrow{\text{d}} Normal(0, Var(\varphi(Y))).$$

Thus, we can estimate the variance ν by taking the sample variance of each influence curve and dividing by the total number of units n. We use this to make 95% confidence intervals for each estimate $(\hat{\Psi} - 1.96^*\sqrt{\nu}, \hat{\Psi} + 1.96^*\sqrt{\nu})$. The influence curves for CARE–IPW and IPTW can be found in the Supplementary Materials B.2.

Evaluation

In the simulations, the estimates, variances, and confidence intervals made by each estimator were compared using bias, Monte Carlo standard error, average standard error, confidence interval coverage, and statistical power. For each simulation $s = 1, \ldots, S$, each estimator made an estimate $\hat{\Psi}_s$, with variance ν_s, and a 95% confidence interval based on the estimate and the variance. Bias is the difference between the average estimate across simulations and the statistical parameter $\frac{1}{S}\sum_{s=1}^{S}\hat{\Psi}_s - \Psi$. Monte Carlo standard error is the standard error in the estimates across simulations $\sqrt{Var(\hat{\Psi}_{1:S})}$. Average standard error is the mean of the estimated standard errors across simulations $\frac{1}{S}\sum_{s=1}^{S}\sqrt{\nu_s}$. Confidence interval coverage is the observed proportion of 95% confidence intervals that covered the statistical parameter Ψ across all simulations. Power is the observed proportion of estimates that rejected the null hypothesis of no intervention effect in simulations where there was an intervention effect. Type I error is the observed proportion of estimates that rejected the null hypothesis of no intervention effect in simulations where the null hypothesis was true.

As a secondary analysis, we were interested in whether forecasts with less error made better estimates of Ψ. Thus, we needed metrics for evaluating the accuracy of the forecasts for unexposed units. We only looked at the forecasts for unexposed units because the forecasts were trained on outcomes for unexposed units. Since some units might have been easier or more difficult to forecast, we compared our forecasts to those of 'median forecasts', which was the median of the past outcomes for each unit.

We evaluated the forecasts for the outcomes in the unexposed group with forecasting mean absolute error and relative mean absolute error. The mean absolute error (MAE) was the average absolute difference between the forecasted value and the observed value on the log scale $MAE_{forecast} = \frac{1}{n}\sum_{i=1}^{n} |\log(Y_{i,T}) - \log(\tilde{Y}_i)|$. We used the log scale because all of the outcomes were estimated using Poisson generalized linear models, which make estimates on the log scale. The relative mean absolute error (rMAE) is the mean absolute error of the forecasting model divided by the mean absolute error of the median forecasts:

$$rMAE = \frac{MAE_{forecast}}{MAE_{median}}. \tag{4.5}$$

If the rMAE was less than one, the model forecast had less error than the median forecast; if the rMAE was greater than one, the model forecast had more error than the median forecast; if the rMAE was equal to one, the model forecast had the same amount of error as the median forecast.

For each simulation, we compared the rMAE of the forecasts to the absolute error and standard error of the effect estimate, as well as whether the confidence interval covered the true statistical parameter Ψ and if the null hypothesis of no exposure effect was correctly rejected. We fit penalized regression models to compare the forecast accuracy by rMAE to these effect estimation metrics, accounting for the number of exposed provinces and the average DHF incidence rate amongst the unexposed provinces.

4.3.2.1 Data availability

All simulations were run using R version 3.4.3.[154] Simulations were run in parallel on 15 cores on a remote server. To maintain reproducibility and make sure that the same samples were drawn for each scenario across simulations, we set a seed for the random number generator for each simulation based on the simulation number. The code used for this project can be found on GitHub.

4.4 Synthetic simulation

Prior to implementing the estimators and forecasts on real data, we conducted a synthetic simulation where we generated the baseline covariates, exposure mechanism, and outcomes to investigate whether forecasts could improve effect estimates in a completely controlled setting.

4.4.1 Setup

4.4.1.1 Data generation

To relate the synthetic simulation to our case study, we generated observations (which we called 'incidence rates') for 76 units ('provinces') at 16 times ('years'). We generated province-level average incidence rates α_i from a Poisson distribution, such that some provinces had naturally higher levels of incidence rates than others. We generated 29 baseline covariates $\mathbf{X}^p_{i,t}$ using a multivariate normal distribution for all provinces and years. The first 9 baseline covariates were generated to be similar to those of temperature, rainfall, and DHF incidence for November, December, and January. Each covariate was highly correlated with the other covariates of the same type (*e.g.* temperature), with the correlation decreasing with difference in time (*e.g.* there was more correlation between November and December temperature than between November and January temperature). The covariates were also correlated to the other covariates of different types to a lesser degree. The correlation matrix

demonstrating the relationships between these 9 baseline covariates is shown in Supplementary Materials C.1. The remaining 20 covariates were random noise. The baseline covariates for each province had an underlying p-length vector of random effects $\boldsymbol{\mu}_i^p$, to represent that some provinces were rainier, warmer, or had higher early season DHF incidence than other provinces on average.

At the exposure time, we allocated interventions from a binomial distribution with the probability determined by an inverse-logit function consisting of two of the baseline covariates $\mathbb{P}(A_i = 1 \mid \mathbf{X}_{i,T}^k)$; we refer to these covariates as 'November temperature' and 'January rainfall' to relate them to the case study, though their names and relationships to the outcome were set arbitrarily. We drew the potential outcome without the exposure $Y(0)_{i,t}$ from a Poisson distribution whose mean was determined by a function of the province-level random effect α_i and the common cause baseline covariates $\mathbf{X}_{i,T}^k$. At the exposure time T, we calculated the potential outcome with the exposure $Y(1)_{i,T}$ as a function of the potential outcome without the exposure and January rainfall, such that the expected effect size would be equal to 20. The exposure A_i determined which potential outcome was observed at the exposure time. We looked at two scenarios, one where the exposure A_i had an effect on $Y(1)_{i,T}$ and another under the null hypothesis of no exposure effect.

$$
\begin{aligned}
\alpha_i &\sim \text{Poisson}(100) \\
\boldsymbol{\mu}_i^p &\sim \text{MVN}(\mathbf{0}, \mathbf{I}) \\
\mathbf{X}_{i,t}^p &\sim \text{MVN}(\boldsymbol{\mu}_i^p, \boldsymbol{\Sigma}) \\
A_i &\sim \text{Binomial}(logit^{-1}(0.75^*\text{Nov temp}_{i,T} + 0.75^*\text{Jan rain}_{i,T})) \\
Y(0)_{i,t} &\sim \text{Poisson}(\alpha_i + 2^*\text{Nov temp}_{i,t} + 10^*\text{Jan rain}_{i,t}) \\
Y(1)_{i,T} &= Y(0)_{i,T}, \text{ under the null} \\
&= Y(0)_{i,T} - 20 - 4^*\text{Jan rain}_{i,T}, \text{ o.w.}
\end{aligned}
$$

We generated 50 sets of baseline covariates and outcomes prior to the exposure time. For each set, we simulated 50 exposure allocations and outcomes at the exposure time for a total of 2,500 simulations.

4.4.1.2 Forecasts

For the synthetic simulation, we used the first 10 years of each simulated dataset as the training phase and the next 5 years as the testing phase. The forward-backward covariate selection process used in the training phase chose baseline covariates to add to a Poisson generalized linear regression model. The model that had the least error in the testing phase was used to make forecasts for the outcome at the exposure time. This process was run for all 50 sets of baseline covariates and outcomes that were generated prior to the exposure time.

4.4.2 Results

4.4.2.1 Effect estimation

τ	Ψ^τ	Estimator	Bias	Monte Carlo SE	Average SE	CI coverage	Power
0	0	Unadj	13.2	4.9	4.9	22.9%	–
		IPTW(G)	5.4	83.1	93.7	94.0%	–
		IPTW(L)	9.1	5.8	21.9	100.0%	–
		IPTW(F)	-0.8	12.2	29.5	100.0%	–
20	-19.5	Unadj	15.4	5.3	5.3	17.1%	12.6%
		IPTW(G)	7.3	80.3	85.7	78.6%	3.4%
		IPTW(L)	12.5	5.8	20.1	99.8%	0.0%
		IPTW(F)	-0.1	11.6	27.4	100.0%	0.0%

Table 4.2. Results for the unadjusted and inverse probability of treatment weighting (IPTW) estimators in the synthetic simulation. There were two effect size scenarios (τ), the null hypothesis of no exposure effect and the alternate with an average effect size of 20. Ψ^τ is the average treatment effect for each effect size, as found by taking the average difference in the potential outcomes.

In each simulation, there were 76 provinces with 38 of them receiving the exposure on average (range: 24-54). The average potential outcome without the exposure was 100.2 across all units (range: 23-210), which was the same for the average potential outcome with the exposure under the null. The average potential outcome with the exposure when the null was false was 80.7 (range: 1-202). The average treatment effect (Ψ^{τ}) was -19.5 when the null was false, as in the exposure reduced the outcome by 19.5 units.

The performance of the effect estimators varied across approaches. For reference, the estimates from the unadjusted estimator were biased in both scenarios, with high bias and low confidence interval coverage (Table 4.2). The IPTW(F) estimator had low bias and high confidence interval coverage in both scenarios. The IPTW(G) estimator had less bias than IPTW(L) or the unadjusted, but large Monte Carlo and average standard errors. The IPTW(L) estimator had high confidence interval coverage in both scenarios, but more bias than any estimator besides the unadjusted estimator. All of the IPTW estimators had low statistical power.

When CARE–IPW used the forecasted values to estimate the propensity scores, it had low bias and high confidence interval coverage in both scenarios, as well as high statistical power when the null was false (Table 4.3). When CARE–IPW used the forecasted values or lasso to estimate the outcomes, the Monte Carlo and average standard errors decreased relative to the IPTW estimator using the same propensity scores. This simulation demonstrates that using forecasted values in effect estimators can improve effect estimates over traditional techniques, especially when the forecasted values are used for estimating propensity scores.

4.4.2.2 Forecasts

There were two covariates associated with the outcome in the synthetic simulation, January rainfall and November temperature, with January rainfall as the stronger

τ	Estimator	Bias	Monte Carlo SE	Average SE	CI coverage	Power
0	CARE(G,G)	-0.0	12.0	11.9	97.4%	–
	CARE(L,L)	3.6	3.4	3.6	84.1%	–
	CARE(F,F)	0.3	4.2	4.1	95.2%	–
	CARE(L,F)	-1.7	4.5	5.4	98.6%	–
	CARE(F,L)	0.3	2.9	3.1	95.9%	–
20	CARE(G,G)	5.0	29.4	16.2	32.6%	64.8%
	CARE(L,L)	8.0	3.7	4.1	48.8%	84.2%
	CARE(F,F)	0.9	4.3	4.8	95.0%	98.2%
	CARE(L,F)	-0.5	5.0	7.2	99.1%	96.4%
	CARE(F,L)	4.5	3.3	3.3	70.0%	99.6%

Table 4.3. Results for the CARE–IPW estimator in the synthetic simulation when using different methods to estimate the outcome and the propensity scores.

of the two. Of the 50 forecasting models used to predict the exposure times in the synthetic simulation, 50 selected January rainfall as a covariate and 43 selected November temperature. Even though 44 forecasting models included other covariates, the relative mean absolute error (rMAE; see Section 4.3.2) of the forecasts for the unexposed provinces at the exposure time was below one for all simulations (mean: 0.54, range: 0.29 to 0.88), indicating that the forecasting model was always better than the median model. This suggests that even if spurious covariates were included in the forecasting model, they did not receive so much weight as to strongly affect the forecasting performance.

Simulations with better forecasting accuracy had better effect estimation performance when using the IPTW(F) and CARE(F,F) estimators. The effect estimates had less absolute error and standard error when there was less forecast rMAE. However, both estimator's absolute error and the standard error were more strongly correlated with the the mean DHF incidence rate than forecasting accuracy. The IPTW(F) estimator was also strongly associated with the number of provinces assigned to the exposure group. Confidence interval coverage and statistical power were not associated

with forecast rMAE for either estimator. Figures can be found in the Supplemental Materials C.2.

4.5 Historical simulation study

We designed a historical simulation study based on data observed prior to 2016 to evaluate the performance of the estimators with real data.

4.5.1 Setup

4.5.1.1 Forecasts

For the historical simulation study, we alternatingly selected each year from 2000 to 2015 to be the exposure time and reorganized the other years into the training and testing phases. For instance, when 2008 was selected to be the exposure time, the years from 2009-2015 and 2000-2002 became the training phase while 2003-2007 became the testing phase. We used the same covariate selection process as in the synthetic simulation, whereby the forecasting model that minimized the testing phase mean absolute error was chosen to make forecasts at the exposure time.

4.5.1.2 Exposure allocations

For each exposure time, we simulated 100 exposure allocations at 4 effect levels τ for a total of 6,400 simulated exposure datasets. The probability of allocation to exposure $\mathbb{P}(A_i = 1)$ was based on a linear function of population, pre-exposure DHF incidence, weather covariates, and historical DHF incidence, similar to the relationships observed for the interventions in 2016, such that there was an average of 10 exposed provinces each year. The exposure effect levels τ included reductions of 5, 10, and 15 cases per 100,000 population, as well as the null hypothesis of no exposure effect. For context, from 2000 to 2015, there was an average of 37.2 (range: 0.15-336) DHF cases per 100,000 population in Thai provinces from July through December, with substantial annual and spatial variability. We added Poisson noise to the originally observed

values $Y_{i,T}$ to simulate the counterfactual outcome without exposure $Y_i(0)$ and applied the exposure effects to determine the counterfactual outcomes with exposure $Y_i(1)$:

$$A_i \sim \text{Binomial}(logit^{-1}(-0.36^*\text{Nov temp}_{i,T} - 0.23^*\text{Jan rain}_{i,T} + 2.4^*log(\text{Population}_{i,T}) - 0.54^*\text{Jan DHF rate}_{i,T} + 0.04^*\text{Median DHF rate}_{i,T}))$$

$$Y_i(0) \sim \text{Poisson}(Y_{i,T})$$

$$Y_i(1) = Y_i(0) - \tau^*\text{Population}_{i,T}/100,000,$$

We found the statistical parameter Ψ^τ by taking the average difference in the counterfactual outcomes for each exposure level τ.

4.5.2 Results

4.5.2.1 Effect estimation

The statistical parameters Ψ^τ and the unadjusted estimates of that parameter $\hat{\Psi}^\tau$ are presented in Table 4.4. Based on the results from the unadjusted estimator, the historical simulation had weaker confounding than the synthetic study (as shown by less bias and greater confidence interval coverage) and there was more variability in the estimates (as shown with greater MC and average standard error). There was also substantial year-to-year variability in bias and confidence interval coverage for the unadjusted estimator (Figure 4.4).

As in the synthetic simulation, the IPTW estimator performed best when using the forecasted values in the propensity score estimates (Table 4.4). The IPTW(G) estimator was very biased with low confidence interval coverage. The IPTW(L) estimator also had more bias and worse confidence interval coverage than the unadjusted estimator. The IPTW(F) estimator had less bias and more confidence interval coverage than the unadjusted estimator, though the confidence interval coverage was still below 95%. The IPTW(F) estimator had less bias than the unadjusted estimator in 64.1% of years and as the least as much confidence interval coverage in 96.9% of years.

τ	Ψ^τ	Estimator	Bias	Monte Carlo SE	Average SE	CI coverage	Power
0	0.0	Unadj	-8.2	12.3	9.4	63.2%	–
		IPTW(G)	-29.9	23.1	12.9	2.8%	–
		IPTW(L)	-18.4	14.1	12.8	48.5%	–
		IPTW(F)	-7.3	11.4	18.4	87.3%	–
5	-4.7	Unadj	-8.2	12.3	9.4	63.0%	52.3%
		IPTW(G)	-26.2	22.0	12.2	7.1%	98.6%
		IPTW(L)	-16.5	13.6	12.0	50.1%	64.3%
		IPTW(F)	-7.0	11.3	17.1	84.8%	25.9%
10	-9.1	Unadj	-7.8	12.2	9.2	63.6%	63.5%
		IPTW(G)	-22.6	21.4	11.6	14.6%	98.9%
		IPTW(L)	-14.5	13.4	11.2	53.4%	74.1%
		IPTW(F)	-6.6	11.1	15.9	81.9%	40.9%
15	-13.0	Unadj	-7.4	12.1	9.0	67.9%	73.2%
		IPTW(G)	-19.4	21.2	11.1	24.5%	99.2%
		IPTW(L)	-12.8	13.3	10.4	59.2%	81.4%
		IPTW(F)	-6.1	11.0	14.7	84.0%	54.6%

Table 4.4. The values of the statistical parameter Ψ^τ for each level of exposure τ and the corresponding estimates by the unadjusted and IPTW estimators.

The CARE–IPW estimators varied in performance by method and effect size (Table 4.5). The CARE(G,G) estimates were all close to zero, regardless of the year or effect size τ. Under the null hypothesis of no effect of exposure, all of the other CARE–IPW estimators had lower bias and higher confidence interval coverage than the unadjusted estimator; and CARE(L,L) and CARE(L,F) had less bias and higher confidence interval coverage than any IPTW estimator. When there was an exposure effect, the CARE(L,F) estimator had the highest confidence interval coverage of any estimator, with all estimates above or near the nominal 95% level. CARE(L,F) had lower bias at higher effect levels ($\tau = 10, 15$) than at lower effect levels. CARE(L,L) and CARE(F,L) had low bias and high confidence interval coverage at low effect levels, but more bias and lower confidence interval coverage at higher effect levels. CARE(F,F) had similar levels of bias across effect levels, with better confidence interval coverage at higher effect levels than at lower effect levels. The CARE(L,F)

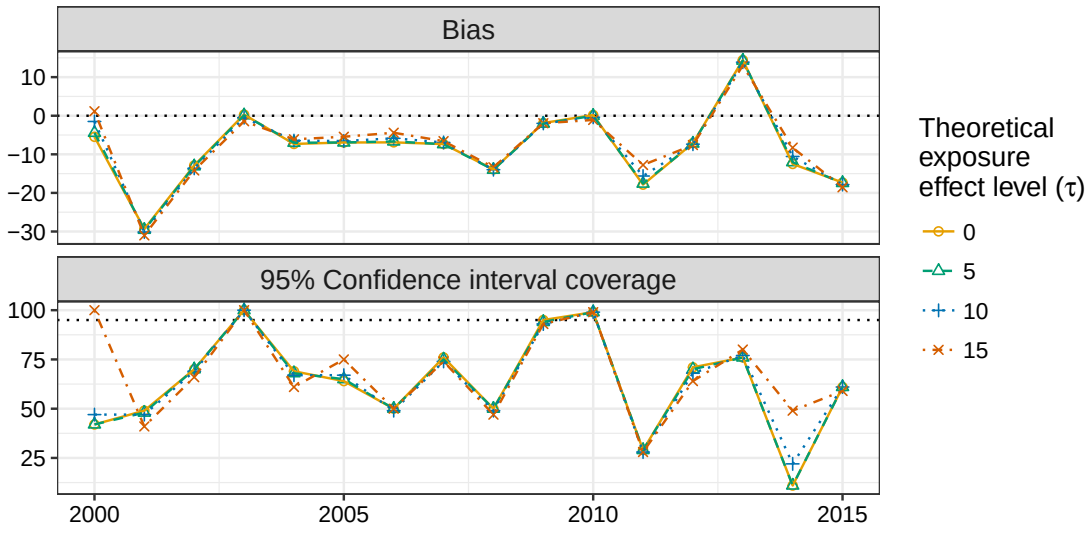

Figure 4.4. The performance of the unadjusted estimator across years by exposure effect size.

estimator made estimates with less bias than the IPTW(F) estimator in 85.9% of years and with greater confidence interval coverage in 57.8% of years.

4.5.2.2 Forecasts

Unlike the synthetic simulation, we do not know the true underlying model for the outcome in the historical simulation. In the 16 forecasting models used in the historical simulation (one for each season), the most commonly chosen covariates were November temperature (13 times), November rainfall (10), and December temperature (8).

The forecasting models performed poorly in predicting DHF incidence in the unexposed provinces at the exposure time relative to the median model. The forecasting rMAE was near one on average (mean: 1.02, range: 0.65 to 1.33) and had more error than the median model 62.6% of the time. Despite the poor forecasting performance, the effect estimators using forecasted values still had better confidence interval coverage

τ	Estimator	Bias	Monte Carlo SE	Average SE	CI coverage	Power
0	CARE(F,L)	-1.8	7.2	7.0	85.4%	–
	CARE(L,L)	-1.8	5.7	6.0	94.6%	–
	CARE(F,F)	-4.3	10.9	10.4	82.9%	–
	CARE(L,F)	-3.2	8.1	8.5	92.6%	–
5	CARE(F,L)	-0.0	7.3	7.2	88.8%	22.0%
	CARE(L,L)	0.3	5.9	6.2	90.1%	17.6%
	CARE(F,F)	-4.0	10.9	10.9	89.4%	23.7%
	CARE(L,F)	-2.0	8.2	8.8	96.2%	18.9%
10	CARE(F,L)	1.7	7.3	7.5	81.1%	31.7%
	CARE(L,L)	2.5	6.2	6.5	79.1%	29.2%
	CARE(F,F)	-3.5	10.8	11.4	93.3%	32.9%
	CARE(L,F)	-0.8	8.3	9.1	96.4%	32.4%
15	CARE(F,L)	3.3	7.4	7.8	75.6%	39.6%
	CARE(L,L)	4.4	6.6	6.7	68.1%	36.5%
	CARE(F,F)	-3.0	10.7	12.0	94.7%	41.8%
	CARE(L,F)	0.4	8.6	9.5	94.4%	40.2%

Table 4.5. Results for the CARE–IPW estimator in the historical simulation when using different methods to estimate the outcome and the propensity scores.

than effect estimators that did not use forecasted values, especially for the IPTW estimators.

When the forecasts had less rMAE, estimates from the IPTW(F) and CARE(L,F) estimators had less absolute error. The IPTW(F) estimator also had higher confidence interval coverage when there was less forecasting error. As in the synthetic simulations, these metrics were more strongly correlated with the number of provinces with the exposure and the average incidence rate in the unexposed provinces than with forecasting rMAE. Figures can be found in the Supplemental Materials C.2.

4.6 Application

4.6.1 Setup

We ran an analysis to determine whether Zika emergency operations centers (EOCs) were associated with a reduction in subsequent DHF incidence in Thailand in 2016. We considered the intervention group to be the 10 provinces that reported Zika cases prior to 29 June 2016.

To generate our forecasts, we used the same model selection procedure as described above, data from 2000 to 2009 as the training phase, and data from 2010 to 2015 as the testing phase. The model that made the forecasts with the least mean absolute error in the testing phase was used to forecast 2016.

We estimated the size of the effect associated with the deployment of the Zika EOCs using a subset of the effect estimators. Along with the unadjusted estimator, we used the IPTW(F) estimator, which was the best performing IPTW estimator; the CARE(F,F) estimator, which had the most statistical power in the synthetic simulations; and the CARE(L,F) estimator, which had the best confidence interval coverage in the historical simulations.

4.6.2 Results

4.6.2.1 Forecasts

The covariate selection procedure chose a forecasting model for 2016 based on November and December temperature as well as January rainfall. The forecasting model predicted that 2016 was going to have more DHF incidence than the median year, as the forecasted values for all 76 provinces were larger than the median DHF incidence rates for each province. Furthermore, there were 10 forecasted outbreaks, defined as years where the DHF incidence rate exceeds the median plus two standard deviations for a province.

In the end, 2016 had less DHF incidence (mean: 25.8, range: 0.5 to 206.8) than expected; only 20 provinces exceeded their median DHF incidence rate, with only 2 experiencing outbreaks. In all, the forecasted values were larger than the observed values in 69 of the 76 provinces; all 10 of the provinces that received the Zika EOCs had forecasted values larger than their observed values (Figure 4.5). The mean absolute error of the forecasted values for the provinces in the unexposed group was 1.55 and the relative mean absolute error was 1.4. In the historic simulations, only the year 2000 had simulations with more mean absolute error and none of the simulations had as much relative mean absolute error. While forecasts with low error are preferable, the effect estimators were able to make accurate estimates even forecasts were poor, in part because performance was more strongly associated with the annual DHF incidence rate in unexposed provinces (Supplemental Materials C.2). In simulations with annual DHF incidence rates around 25 cases per 100,000 population, the IPTW(F) and CARE(L,F) estimators had lower than average absolute error and standard error and above average confidence interval coverage.

Lasso regressions were used to estimate the outcomes and propensity scores for several of the CARE–IPW estimators. The lasso regression for the CARE–IPW outcomes had four covariates: 2010 DHF incidence rate, December rainfall, January rainfall, and maximum DHF incidence rate since 2000.

4.6.2.2 Effect estimation

In our analysis, each of the estimators suggested that Zika interventions reduced subsequent DHF incidence to varying degrees (Figure 4.6). The unadjusted, IPTW(F), and CARE(F,F) estimators made similar estimates (with similarly large uncertainty) of the difference in DHF incidence rate associated with the Zika EOCs, with each point estimate near -15 cases per 100,000 population. Their similarity to the unadjusted estimator suggests that the IPTW(F) and CARE(F,F) estimators did not account for

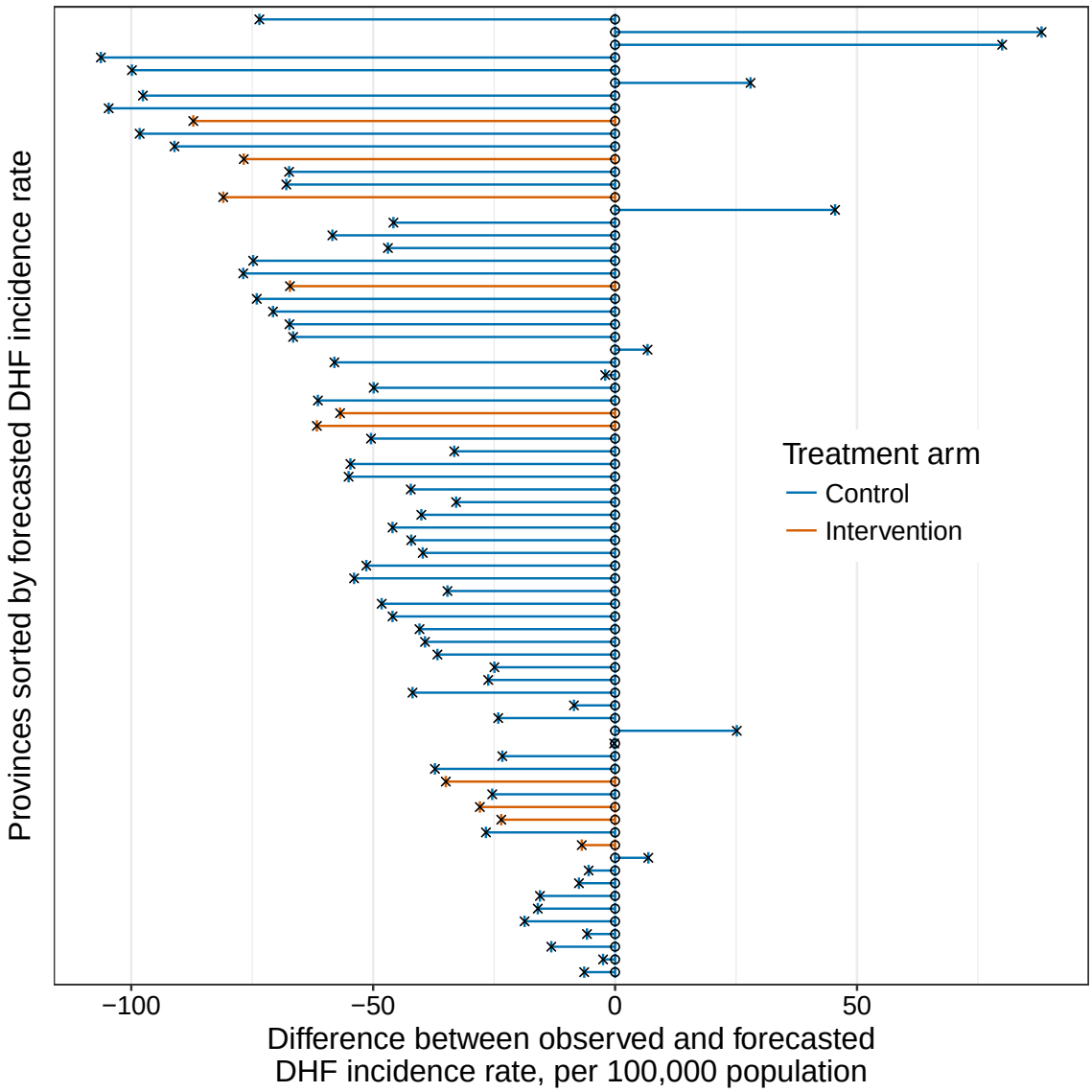

Figure 4.5. The difference between observations and forecasts of the dengue hemorrhagic fever (DHF) incidence rate in each province. The circles represent forecasted values and the x's represent observed values. Blue lines indicate that the province is in the control group, while orange lines indicate that the province is in the intervention group, having reported a Zika case prior to 29 June 2016.

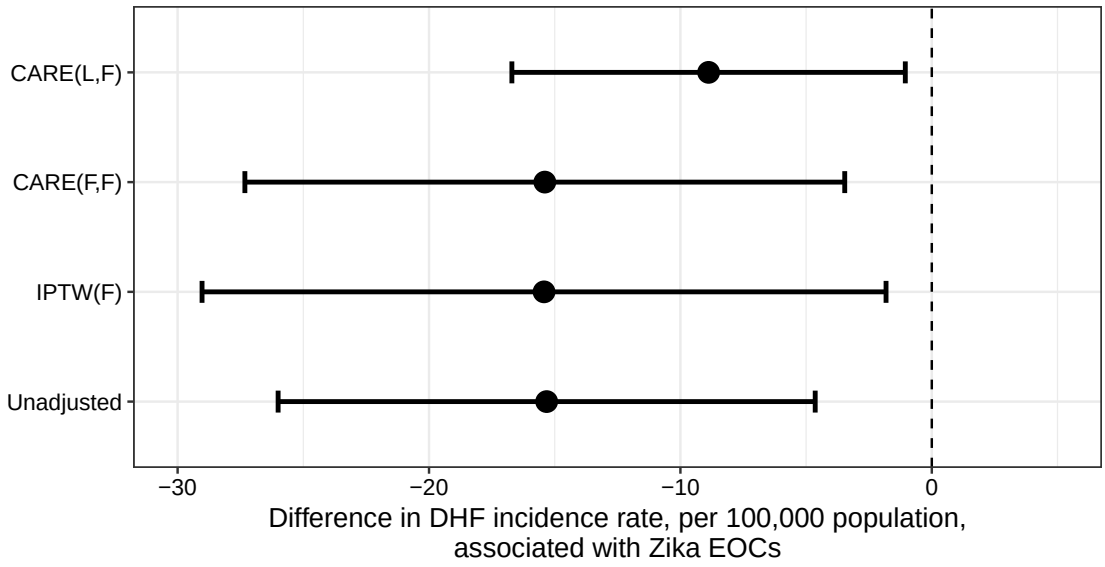

Figure 4.6. The estimates and 95% confidence intervals for the intervention effect made by selected estimators.

much confounding between the exposure and the outcome. The CARE(L,F) estimator estimated that there was a reduction of about 8.9 cases per 100,000 population with less uncertainty.

4.7 Discussion

In this paper, we demonstrated that forecasts can be used to improve estimates of the effect of an intervention. In a synthetic simulation, where we could control all of the causes of the exposure and outcome, using forecasted values in effect estimation reduced bias and increased confidence interval coverage when using the inverse-probability of treatment weighting (IPTW) estimator and covariate-adjusted residuals estimator with inverse-probability weighting (CARE–IPW). In a historical simulation using Thailand dengue hemorrhagic fever (DHF) data, which had less confounding but more variability than the synthetic simulation, using forecasted values improved confidence interval coverage over using lasso regression. In both simulations, when forecasts had

less prediction error the effect estimators had less error; in the historical simulation, forecasts with less prediction error also had greater confidence interval coverage.

Forecasting exhibited utility as a means of covariate selection and dimension reduction, especially for propensity scores. The IPTW and CARE–IPW estimators rely on propensity scores that account for common causes, but can be overfit to non-confounders or random noise. By forecasting outcomes, only covariates that affect the outcome are selected; and covariates that affect only the exposure, but not the outcome, are not selected. By using a rigorous cross-validation and testing framework, only the covariates with the strongest associations with the outcome are included. Thus, propensity scores are fit only to the potentially-confounding covariates and are unlikely to be overfit. That the IPTW and CARE–IPW effect estimates using forecasted values performed better than those using lasso regression in the synthetic simulation is an encouraging outcome. Lasso regression is a commonly-used method for experiments with small sample sizes and large numbers of covariates.[78] However, even lasso regression may be prone to picking up false signals from random noise or collinear covariates with few observations.[121, 201] The forecasted outcomes avoided these pitfalls due, at least in part, to the fact that the forecasting procedure was able to use a larger sample of observations. Also of note is that the CARE–IPW estimator performed best with a mixture of forecasted values and lasso regression. Further investigation is required to determine which situations benefit most from the use of forecasted values and how best to implement them.

Determining whether more accurate forecasts improve effect estimation is another area for future work. In this paper, we restricted the forecasts to use only generalized linear regression in order to isolate the improvement from using forecasted values for covariate selection and dimension reduction, as opposed to the improvement from using more advanced forecasting techniques. In the simulations, the effect estimation error was more strongly associated with the incidence rates of the observations than

with the accuracy of the forecasts. A focused study could provide more evidence for the relationship between forecast accuracy on effect estimation.

We used the effect estimators with forecasted values to estimate whether emergency operations centers (EOCs), deployed in Thailand during the 2016 Zika pandemic, were associated with a reduction in subsequent DHF incidence. The CARE(L,F) estimate showed a reduction of 8.9 (95% CI: 1.1, 16.7) DHF cases per 100,000 population associated with the Zika EOCs. For this estimate to correctly identify the average treatment effect of the Zika EOCs on DHF incidence, we assume that each province had a non-zero probability of receiving the intervention and that there are no unmeasured confounding covariates. We expect that every province had a non-zero probability of reporting a Zika case and receiving a Zika EOC. One potential unmeasured confounding covariate is the population susceptibility to dengue in each province. There is some evidence of antibody-dependent enhancement between Zika and dengue,[38, 186] which would mean that past dengue incidence could influence the number of Zika cases in a province. While we included DHF incidence from past seasons, a better dengue susceptibility metric may improve effect estimation. Additionally, due to the restriction to generalized linear models, the forecasting and estimation models may be missing non-linear relationships between covariates and the exposure and outcome. Using non-parametric techniques for forecasting and for estimating outcomes and propensity scores is an area for future work.

The methods proposed in this paper suggest that forecasts could be utilized in other effect estimation contexts. We demonstrated that forecasts can be incorporated into effect estimates under the following conditions: observed outcomes for a unit are independent of the outcomes and exposures for other units; exposures are binary and occur at a single time point; baseline covariates precede the exposure which precedes the outcome; and that observations occur at multiple time points for the purpose of training forecasts. While we specifically considered data where the baseline covariates,

exposure, and outcome all occurred sequentially within a single time point, these methods could be extended to settings where data are observed across multiple time points. For instance, consider a situation in which researchers make weekly influenza incidence forecasts based on previously observed influenza and weather data for a number of school districts. Near the peak of the influenza season, the researchers share their forecasts with school administrators and advise them to close schools for a week to reduce influenza transmission. The school administrators (independently) use their discretion in determining school closures, with higher risk school districts being more likely to close. The reseachers could use the forecasts made prior to the school closure week for the influenza incidence occurring after the school closure week in their effect estimation process. While the relative efficiency of using forecasts against existing methods has yet to be evaluated for this scenario, the promising results presented in this paper warrant the exploration of incorporating forecasts into other effect estimation settings.

APPENDIX A

PROSPECTIVE FORECASTS OF ANNUAL DENGUE HEMORRHAGIC FEVER INCIDENCE IN THAILAND, 2010-2014 SUPPLEMENT

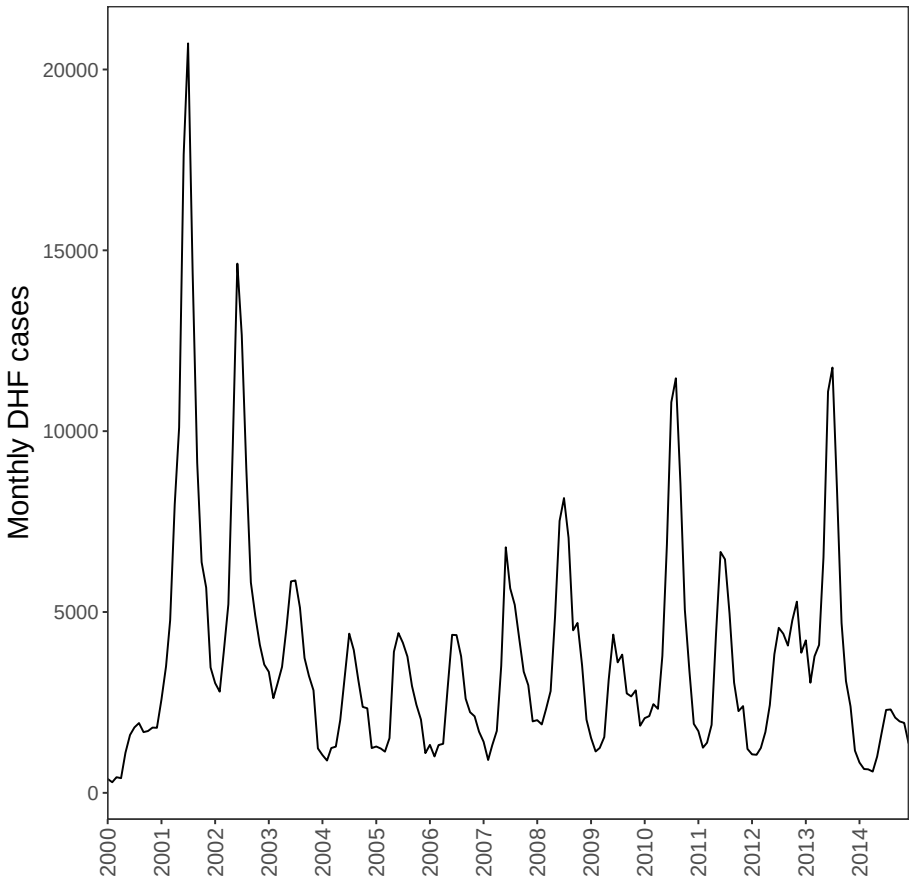

Figure A.1. Aggregated time series of dengue hemorrhagic fever cases from 2000-2014.

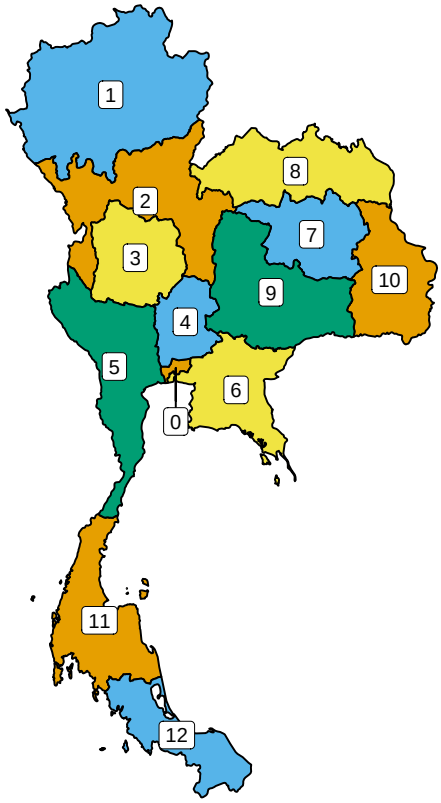

Figure A.2. Map of the Thailand Ministry of Public Health administrative regions (MOPH regions). These 13 MOPH regions are geographically clustered sets of 4-8 provinces (with the exception of Bangkok, region 0, which is its own region) co-operatively managed by a regional health office.

Covariate Type	Covariate Name	Incidence-only	WIP
Incidence	estimated relative susceptibility rate		
	last high-season incidence rate		
	last post-season incidence rate		
	pre-season incidence rate	✓	✓
Demographics	population per square kilometer		
	provincial population		✓
Humidity	maximum low-season humidity		
	minimum low-season humidity		
	mean January humidity		
	mean February humidity		
	mean March humidity		
Rainfall	maximum low-season rainfall (NOAA)		
	total low-season rainfall (ESRL)		
	total low-season rainfall (NOAA)		
	maximum January rainfall (NOAA)		
	total January rainfall (ESRL)		✓
	total January rainfall (NOAA)		
	maximum February rainfall (NOAA)		
	total February rainfall (ESRL)		
	total February rainfall (NOAA)		
	maximum March rainfall (NOAA)		
	total March rainfall (ESRL)		
	total March rainfall (NOAA)		
Temperature	maximum low-season temperature (NCDC)		
	mean low-season temperature (ESRL)		✓
	minimum low-season temperature (NCDC)		
	mean January temperature (ESRL)		✓
	mean January temperature (NCDC)		
	mean January temperature (NOAA)		
	mean February temperature (ESRL)		
	mean February temperature (NCDC)		
	mean February temperature (NOAA)		
	mean March temperature (ESRL)		
	mean March temperature (NCDC)		
	mean March temperature (NOAA)		

Table A.1. Covariates considered for inclusion prior to model selection. "Incidence-only" indicates the covariates that were included in the incidence-only model. "WIP" indicates the covariates that were included in the weather, incidence, and population model.

Model	MAE	rMAE	% of forecasts better than baseline	80% PI coverage	AIC
WIP	0.64	0.87	56.3	69.7	**9909**
Incidence-only	**0.59**	**0.81**	**64.7**	**80.0**	10055
Baseline	0.73	1.00			

Table A.2. Results for each model across all regions and years in the testing phase. Numbers in bold highlight which model performed best for each metric.

Model	Year	Mean provincial incidence	Outbreaks	MAE	rMAE	% of forecasts better than baseline	80% PI coverage
WIP				0.53	0.73	60.5	76.3
Incidence-only	2010	78	12	**0.53**	**0.73**	**68.4**	**80.3**
Baseline				0.72	1.00		
WIP				0.65	1.05	47.4	61.8
Incidence-only	2011	47	2	**0.59**	**0.95**	**55.3**	**78.9**
Baseline				0.61	1.00		
WIP				0.43	0.90	60.5	**86.8**
Incidence-only	2012	48	1	**0.43**	**0.90**	**61.8**	89.5
Baseline				0.48	1.00		
WIP				0.57	0.79	51.3	**75.0**
Incidence-only	2013	74	23	**0.56**	**0.77**	**55.3**	85.5
Baseline				0.73	1.00		
WIP				1.00	0.89	61.8	48.7
Incidence-only	2014	24	0	**0.85**	**0.75**	**82.9**	**65.8**
Baseline				1.13	1.00		

Table A.3. Annual results for each model across all regions in the testing phase. Numbers in bold highlight which model performed best for each metric in each year.

As displayed in Table A.4, the smallest mean absolute error (MAE) by any model for any region was for Bangkok (MOPH region 0) using the incidence-only model (MAE=0.286). However, because the baseline MAE for Bangkok was only slightly higher (MAE=0.289), the incidence-only model relative MAE (rMAE) was the second-largest of any region (rMAE=0.99). Thus, even though the incidence-only model accurately forecasted DHF incidence in Bangkok, it didn't add much value over a

ten-year median, due in part to there being no outbreaks in Bangkok during the testing phase.

Conversely, the incidence-only model had about twice as much error in MOPH region 12 (MAE=0.59) as in Bangkok. However, the baseline model had nearly three times as much error than in Bangkok (MAE=0.86), so the incidence-only rMAE for MOPH region 12 was the lowest of any model for any region (rMAE=0.69). Thus, despite greater absolute error from the incidence-only model forecasts, there was more added benefit for that region over the baseline forecasts than for Bangkok. These examples demonstrate how MAE and rMAE can be used in tandem to give a more complete evaluation of model performance.

Model	MOPH Region	# of Provs	Mean provincial incidence	Outbreaks	MAE	rMAE	% of forecasts better than baseline	80% PI coverage
WIP					0.62	0.72	68.6	65.7
Incidence	12	7	83	3	**0.59**	**0.69**	**77.1**	**74.3**
Baseline					0.86	1.00		
WIP					0.58	0.84	70.0	70.0
Incidence	9	4	74	3	**0.48**	**0.70**	**75.0**	**85.0**
Baseline					0.69	1.00		
WIP					**0.78**	**0.71**	70.0	66.7
Incidence	8	6	37	5	0.83	0.75	70.0	66.7
Baseline					1.10	1.00		
WIP					0.56	0.82	60.0	**76.0**
Incidence	10	5	43	4	**0.49**	**0.71**	**72.0**	88.0
Baseline					0.69	1.00		
WIP					**0.82**	**0.74**	77.5	50.0
Incidence	1	8	45	12	0.94	0.84	77.5	**55.0**
Baseline					1.11	1.00		
WIP					0.56	0.76	60.0	**75.0**
Incidence	7	4	51	4	**0.55**	**0.75**	**75.0**	70.0
Baseline					0.74	1.00		
WIP					0.71	0.79	62.9	62.9
Incidence	11	7	72	3	**0.69**	**0.77**	**65.7**	**77.1**
Baseline					0.90	1.00		
WIP					0.51	1.01	42.5	**75.0**
Incidence	6	8	65	1	**0.39**	**0.79**	**55.0**	95.0
Baseline					0.50	1.00		
WIP					0.62	0.93	48.0	72.0
Incidence	2	5	52	2	**0.56**	**0.84**	**68.0**	**76.0**
Baseline					0.66	1.00		
WIP					0.60	1.04	42.5	**80.0**
Incidence	4	8	33	0	**0.52**	**0.88**	**57.5**	92.5
Baseline					0.58	1.00		
WIP					0.68	1.17	52.0	64.0
Incidence	3	5	51	0	**0.57**	**0.98**	**56.0**	**80.0**
Baseline					0.58	1.00		
WIP					0.58	2.00	20.0	100.0
Incidence	0	1	63	0	**0.29**	**0.99**	**40.0**	100.0
Baseline					0.29	1.00		
WIP					0.56	1.24	37.5	**77.5**
Incidence	5	8	47	1	0.47	1.03	**45.0**	92.5
Baseline					**0.46**	**1.00**		

Table A.4. Regional results for each model across all years in the testing phase. Numbers in bold highlight which model performed best for each metric in each region. The regions are sorted by best model performance using relative mean absolute error (rMAE) from lowest to highest.

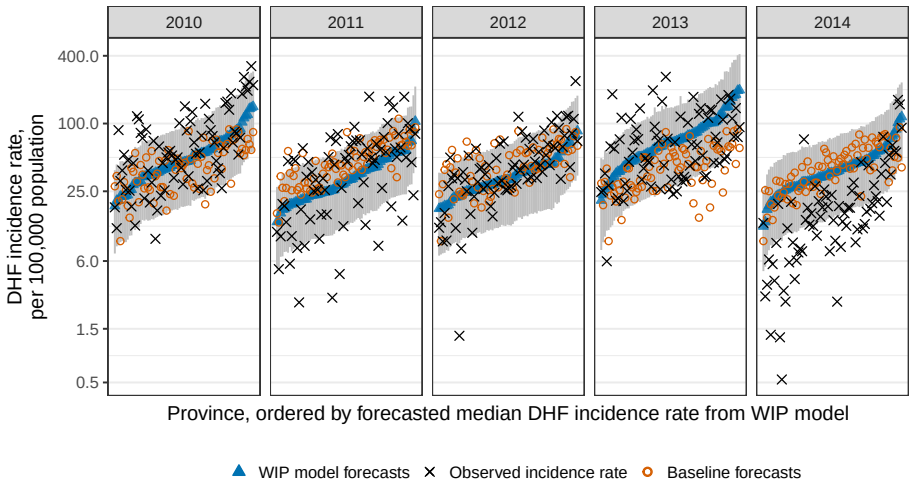

Figure A.3. Weather, incidence, and population (WIP) model forecasts for each year of the testing phase compared to the baseline forecasts and the observed values. Forecasts for the annual dengue hemorrhagic fever (DHF) incidence rate, per 100,000 population, from the WIP model (blue triangles with gray 80% prediction intervals), baseline forecasts (red circles), and observed values (black x's) for each province and year in the testing phase.

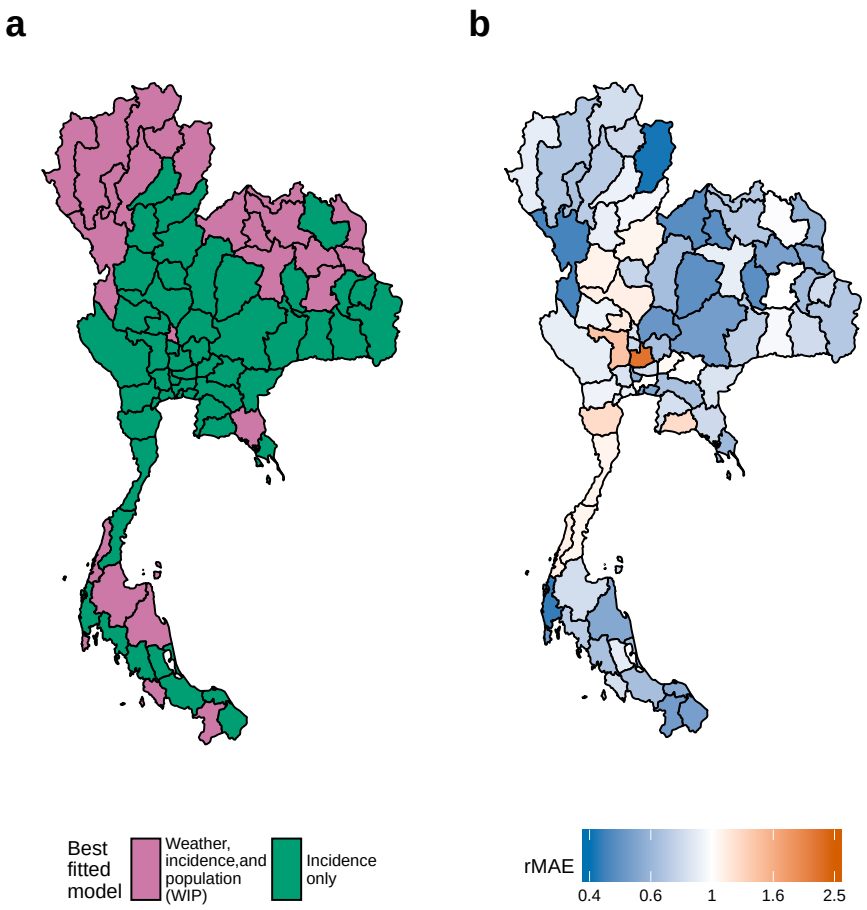

Figure A.4. Geographic variation in model and performance by province. **(a)** The best fitted model in the testing phase for each Thai province, which shows spatial patterns of performance. **(b)** The relative mean absolute error of the forecasts for each province from the models in (a) over the baseline forecasts. Provinces with: less error than the baseline are blue, more error than the baseline are red, and equal to the baseline are white.

The receiver operating characteristic (ROC) curves for the incidence-only and WIP model outbreak forecasts for the testing phase are both significantly above the line of no-discrimination, but are not significantly different from each other (Figure A.5). The incidence-only model area under the ROC curve (AUC; Estimate: 84.2%, 95%CI: 78.5-89.9%) was slightly larger than that of the WIP model AUC (82.9%, 76.3-89.6%). The sensitivity of the WIP model is marginally larger than that of the incidence-only

model when specificity is large, suggesting that the WIP model showed very slightly better performance than the incidence-only model at larger outbreak thresholds.

The predictive distributions samples used to make the outbreak forecasts could have been obtained by estimating parameters in a Bayesian framework, including drawing posterior samples of the dispersion parameter, which may have changed the predictive performance of the models. However, due to the coverage rates observed by our model (80% of forecasts covered by the 80% prediction interval), we did not believe it would be worth the additional computational complexity to use these methods.

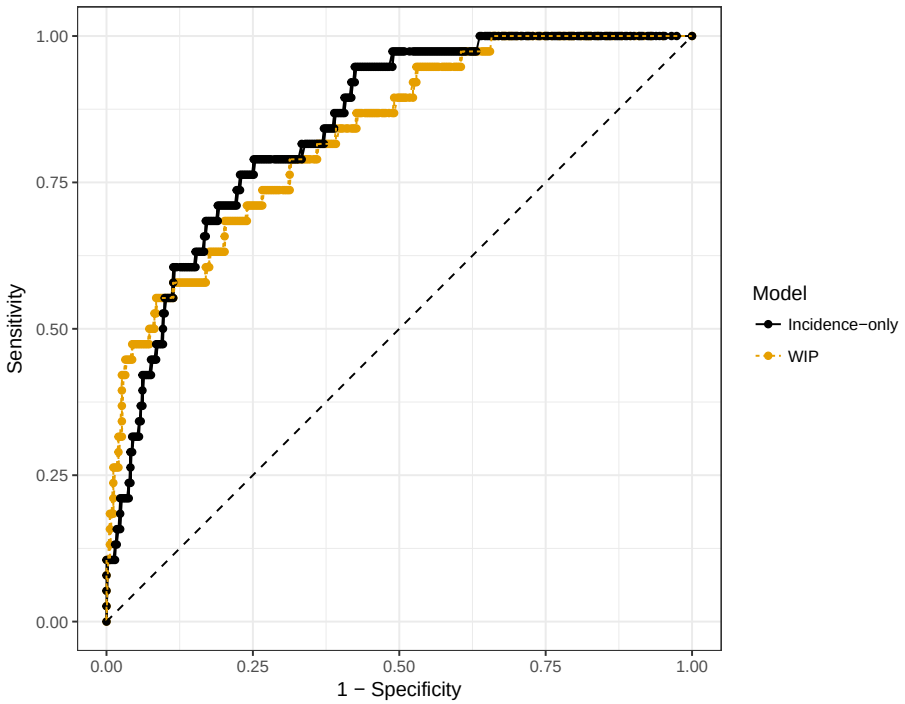

Figure A.5. Comparison of receiver operating characteristic (ROC) curves by model. The ROC curve based on the incidence-only model and weather, incidence, and population (WIP) models' sensitivity and specificity on outbreak forecasts during the testing phase. Both curves are comfortably above the line of no-discrimination (dashed), indicating that their outbreak forecasts are better than random. The AUC for the WIP model (82.9%) is a bit lower than that of the incidence-only model (84.2%).

APPENDIX B

THE COVARIATE-ADJUSTED RESIDUALS ESTIMATOR AND ITS USE IN BOTH RANDOMIZED TRIALS AND OBSERVATIONAL SETTINGS APPENDIX AND SUPPLEMENTAL MATERIALS

B.1 Proofs

B.1.1 CARE in randomized trials

Theorem 1: In a randomized trial, the expectation of the covariate adjusted residuals estimator is equivalent to the expectation of the unadjusted estimator and is thus consistent for the target statistical parameter Ψ.

Proof: We can split the CARE equation (3.7) into an unadjusted component and a predicted component as such:

$$\begin{aligned}\hat{\Psi}^{CARE} = &\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1)} - \frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0)}\right)Y_i \\ &- \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1)} - \frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0)}\right)\hat{\mathbb{E}}(Y_i \mid W_i^Y)\end{aligned} \tag{B.1}$$

The unadjusted component is equivalent to the unadjusted estimator (3.4), which is consistent for the statistical parameter Ψ^{RCT} (see Supplementary Materials B.3.2). Using $\hat{Y} \equiv \hat{\mathbb{E}}(Y \mid W^Y)$, we can find the expectation of the predicted component:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1)}\hat{Y}\right] - \mathbb{E}\left[\frac{\mathbb{I}(A=0)}{\hat{\mathbb{P}}(A=0)}\hat{Y}\right] \\
&= \sum_{w^Y,a,\hat{y}} \frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1)}\hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid A=a, W^Y=w^Y)\mathbb{P}(A=a \mid W^Y=w^Y)\mathbb{P}(W^Y=w^Y) \\
&\quad - \sum_{w^Y,a,\hat{y}} \frac{\mathbb{I}(A=0)}{\hat{\mathbb{P}}(A=0)}\hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid A=a, W^Y=w^Y)\mathbb{P}(A=a \mid W^Y=w^Y)\mathbb{P}(W^Y=w^Y) \\
&= \sum_{w^Y,\hat{y}} \hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid W^Y=w^Y)\mathbb{P}(W^Y=w^Y) - \sum_{w^Y,\hat{y}} \hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid W^Y=w^Y)\mathbb{P}(W^Y=w^Y) \\
&= 0,
\end{aligned}
$$

where summations generalize to integrals for continuous random variables. In the second equality, we used that in a randomized trial $\mathbb{P}(A=a \mid W^Y=w^Y) = \mathbb{P}(A=a)$ and that the predictions $\hat{Y}$ are independent of the exposure A.

Therefore, the expectation of CARE is equivalent to the expectation of the unadjusted estimator and is thus is consistent for the statistical parameter Ψ^{RCT}, which identifies the average treatment effect in randomized trials, where identifiability assumptions hold by design.

Corollary 1.1: If $\hat{Y}$ is a constant (*e.g.* 0 or the mean of all Y), CARE is equivalent to the unadjusted estimator.

Proof: As stated previously, the unadjusted component of CARE (B.1) is equivalent to the unadjusted estimator. If $\hat{Y}$ is equal to a constant C, then the predicted component of CARE is equivalent to:

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i=1)}{\hat{\mathbb{P}}(A=1)} - \frac{\mathbb{I}(A_i=0)}{\hat{\mathbb{P}}(A=0)}\right)C &= \frac{1}{n_1}\sum_{i\in A_i=1} C - \frac{1}{n_0}\sum_{i\in A_i=0} C \\
&= \frac{n_1}{n_1}C - \frac{n_0}{n_0}C \\
&= 0.
\end{aligned}
$$

Thus CARE is equivalent to the unadjusted estimator when the predicted component is equal to any constant.

B.1.2 CARE in observational studies

Theorem 2: In observational settings, where the allocation process is a function of covariates, the covariate-adjusted residuals estimator will be biased for the statistical parameter Ψ.

Proof: The unadjusted component of CARE is equivalent to the unadjusted estimator, which is biased for the statistical parameter Ψ because the outcome and allocation to exposure are no longer independent (Supplemental Materials B.3.3). Furthermore, the expectation of the predicted component of CARE is no longer zero because the probability of exposure is dependent on confounding covariates $\mathbb{P}(A = a \mid W^C = w^C) \neq \mathbb{P}(A = a)$:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1)}\hat{Y}\right] - \mathbb{E}\left[\frac{\mathbb{I}(A=0)}{\hat{\mathbb{P}}(A=0)}\hat{Y}\right] \\
&= \sum_{w^C,\hat{y}} \frac{\mathbb{P}(A=1 \mid W^C = w^C)}{\hat{\mathbb{P}}(A=1)}\hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid W^C = w^C)\mathbb{P}(W^C = w^C) \\
&\quad - \sum_{w^C,\hat{y}} \frac{\mathbb{P}(A=0 \mid W^C = w^C)}{\hat{\mathbb{P}}(A=0)}\hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid W^C = w^C)\mathbb{P}(W^C = w^C) \\
&\neq 0,
\end{aligned}
$$

where summations generalize to integrals for continuous random variables and the predictions $\hat{Y}$ are still independent of the exposure A. This implies that, CARE is not consistent for the statistical parameter Ψ, even if the covariates W^C are sufficient to control for confounding in the predictions for the outcome Y.

Theorem 3: In observational settings, under the strong null hypothesis that there is no exposure effect for all units, the covariate-adjusted residuals estimator will be unbiased for the statistical parameter Ψ if the predictions of the outcome are consistent for the outcome.

Proof: Under the strong null hypothesis, the conditional expectation of the outcome is the same with or without the allocation to exposure ($\mathbb{E}(Y \mid A, W^C) = \mathbb{E}(Y \mid W^C)$). If the estimates of the outcome $\hat{\mathbb{E}}(Y \mid W)$ are consistent for the outcome Y, as is the case in linear regression, then the expectation of CARE is zero, which is the statistical parameter Ψ under the null:

$$\begin{aligned}
\mathbb{E}(\hat{\Psi}^{CARE}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)}\right)\left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C)\right)\right) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)}\right)(Y_i - Y_i)\right) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)}\right)0\right) \\
&= 0.
\end{aligned}$$

B.1.3 CARE–IPW

Theorem 4: The expectation of the covariate-adjusted residuals estimator with inverse probability weighting (CARE–IPW) is equivalent to the expectation of the inverse probability of treatment weighting (IPTW) estimator. Thus, CARE–IPW is consistent for the statistical parameter Ψ in randomized or observational settings.

Proof: The IPTW term of CARE–IPW (3.9) is equivalent to the IPTW estimator (3.5), which is consistent for Ψ when propensity scores $\hat{\mathbb{P}}(A = 1 \mid W^C)$ are consistent for the true conditional probability of exposure $\mathbb{P}(A = 1 \mid W^C)$. The expectation of the predicted term of CARE–IPW is:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1 \mid W^C)}\hat{Y}\right] - \mathbb{E}\left[\frac{\mathbb{I}(A=0)}{1-\hat{\mathbb{P}}(A=1 \mid W^C)}\hat{Y}\right] \\
&= \sum_{w^C,a,\hat{y}} \frac{\mathbb{I}(A=1)\hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid A=a, W^C=w^C)\mathbb{P}(A=a \mid W^C=w^C)\mathbb{P}(W^C=w^C)}{\hat{\mathbb{P}}(A=1 \mid W^C=w^C)} \\
&\quad - \sum_{w^C,a,\hat{y}} \frac{\mathbb{I}(A=0)\hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid A=a, W^C=w^C)\mathbb{P}(A=a \mid W^C=w^C)\mathbb{P}(W^C=w^C)}{1-\hat{\mathbb{P}}(A=1 \mid W^C=w^C)} \\
&= \sum_{w^C,\hat{y}} \hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid W^C=w^C)\mathbb{P}(W^C=w^C) - \sum_{w^C,y} \hat{y}\mathbb{P}(\hat{Y}=\hat{y} \mid W^C=w^C)\mathbb{P}(W^C=w^C) \\
&= 0,
\end{aligned}
$$

where the summations generalize to integrals for continuous variables and the predictions $\hat{Y}$ are independent of the exposure A. Thus, the expectation of CARE-IPW is equivalent to the expectation of the IPTW estimator and is consistent for the statistical parameter Ψ, which is consistent for the ATE under the same identifiability assumptions as for the IPTW estimator.

B.2 CARE–IPW is asymptotically normal

In this Appendix, we discuss the theoretical properties of CARE and CARE–IPW in randomized trials and observational settings. Before doing so, we first review the use of augmented inverse probability weighting for estimation and inference of the G-computation identifiability result in a point-treatment setting.

B.2.1 Review of augmented inverse probability weighting

Suppose our data consist of n independent, identically distributed observations of $O = (W^C, A, Y)$ where W^C are the baseline confounders, A is the exposure, and Y is the outcome. These data are distributed according to some unknown probability distribution $\mathbb{P}$, which is an element of a non-parametric or semi-parametric statistical model $\mathcal{M}$. Non-parametric statistical models place no restrictions of the set of possible

observed data distributions, while semi-parametric statistical models place some restrictions, such as randomization of a binary exposure: $\mathbb{P}(A = 1 \mid W^C) = 0.5$.

We focus on estimation of the G-computation identifiability result (3.2) which would equal the average treatment effect under the randomization and positivity assumptions.[162] The *efficient influence curve* establishes the asymptotic bound for the variance for all regular, asymptotically linear estimators [14] and is given by the following for the statistical estimand $\Psi(\mathbb{P}) = \psi$:[165, 76, 196]

$$\begin{aligned} D^*(\mathbb{P}) = &\left(\frac{\mathbb{I}(A = 1)}{\mathbb{P}(A = 1 \mid W^C)} - \frac{\mathbb{I}(A = 0)}{\mathbb{P}(A = 0 \mid W^C)} \right) \left(Y - \mathbb{E}(Y \mid A, W^C) \right) \\ &+ \mathbb{E}(Y \mid A = 1, W^C) - \mathbb{E}(Y \mid A = 0, W^C) - \psi. \end{aligned}$$

We refer the reader to Kennedy (2017) for an introduction to semi-parametric, efficiency theory.[103]

The augmented inverse probability of treatment weighting (AIPW) estimator directly solves the estimating equation corresponding to this efficient influence curve.[165, 196] Specifically, AIPW is the solution in ψ to the estimating equation

$$\begin{aligned} 0 = &\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0 \mid W_i^C)} \right) \left(Y_i - \hat{\mathbb{E}}(Y_i \mid A_i, W_i^C) \right) \\ &+ \hat{\mathbb{E}}(Y_i \mid A_i = 1, W_i^C) - \hat{\mathbb{E}}(Y_i \mid A_i = 0, W_i^C) - \psi. \end{aligned}$$

Thus, AIPW requires estimation of both the conditional mean outcome $\mathbb{E}(Y \mid A, W^C)$ and the propensity score $\mathbb{P}(A = 1 \mid W^C)$. In most settings, these are unknown quantities and thus considered *nuisance parameters*. However, AIPW is *double robust* in that it will be consistent for Ψ if either nuisance parameter is consistently estimated. Under regularity conditions on the nuisance parameter estimation, AIPW is also asymptotically linear and can be written as an empirical average of a mean-zero, finite-

variance function of the observed data, called the *influence curve*, and a remainder term converging to zero in probability:

$$AIPW - \psi = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_{\mathbb{P}}(1/\sqrt{n})$$

where $\mathbb{E}[IC(O)] = 0$ and $Var[IC(O)]$ is finite. Thus, the central limit theorem applies and AIPW is asymptotically normal. This also provides a straightforward approach to variance estimation; specifically, we estimate AIPW's variance with the sample variance of this estimated influence curve divided by sample size n:

$$\begin{aligned}\hat{IC}^{AIPW} = &\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0 \mid W_i^C)}\right)\left(Y_i - \hat{\mathbb{E}}(Y_i \mid A_i, W_i^C)\right) \\ &+ \hat{\mathbb{E}}(Y_i \mid A_i = 1, W_i^C) - \hat{\mathbb{E}}(Y_i \mid A_i = 0, W_i^C) - \hat{\psi}.\end{aligned}$$

Finally, if both nuisance parameters are estimated at fast enough rates, AIPW is *locally efficient* in that its influence curve equals the efficient influence curve ($IC^{AIPW} = D^*(\mathbb{P})$) and it achieves the smallest possible variance.

B.2.2 Consistency and normality of CARE–IPW

CARE–IPW can be considered a special case of AIPW where the conditional mean outcome is estimated ignoring the exposure: $\hat{\mathbb{E}}(Y \mid A, W^C) = \hat{\mathbb{E}}(Y \mid W^C)$. Thereby, CARE–IPW is the solution in ψ to the estimating equation:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0 \mid W_i^C)}\right)\left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C)\right) - \psi.$$

As a result, CARE–IPW inherits many of the properties of AIPW. First, CARE–IPW will be consistent for $\Psi(\mathbb{P})$ if either the estimated propensity score $\hat{\mathbb{P}}(A = 1 \mid W^C)$ converges to the true propensity score $\mathbb{P}(A = 1 \mid W^C)$ *or* if the predicted outcome in the absence of the exposure $\hat{\mathbb{E}}(Y \mid W^C)$ converges to the true conditional mean

outcome $\mathbb{E}(Y \mid A, W^C)$. This implies that CARE–IPW is only double robust under the null when we have $\mathbb{E}(Y \mid A, W^C) = \mathbb{E}(Y \mid W^C)$. When there is an effect (*i.e.* when the null is false), CARE–IPW relies fully on consistent estimation of the propensity score $\mathbb{P}(A = 1 \mid W^C)$. Second, under the same regularity conditions, CARE–IPW is asymptotically normal, and its variance can be estimated by the sample variance of the following influence curve divided by sample size n:

$$\hat{IC}^{CARE-IPW} = \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0 \mid W_i^C)} \right) \left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C) \right) - \hat{\psi}.$$

Finally, under consistent estimation of both nuisance parameters, which again can only occur under the null, CARE–IPW is locally efficient. In other settings, we do, however, expect CARE–IPW to provide efficiency gains over the IPTW estimator, which is the solution in ψ to the estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1 \mid W_i^C)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0 \mid W_i^C)} \right) Y_i - \psi.$$

Specifically, predicting the outcome with the covariates $\hat{\mathbb{E}}(Y \mid W^C)$ in CARE–IPW should result in a more precise estimator than predicting the outcome with zero as with the IPTW estimator.

B.2.3 Consistency and normality of CARE

CARE is also a special case of AIPW where the propensity score is estimated ignoring the covariates $\hat{\mathbb{P}}(A = 1)$ and the conditional mean outcome is estimated ignoring the exposure $\hat{\mathbb{E}}(Y \mid A, W^C) = \hat{\mathbb{E}}(Y \mid W^C)$. In other words, CARE is the solution in ψ to the estimating equation:

$$0 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0)} \right) \left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C) \right) - \psi.$$

As a result, CARE will be consistent for $\Psi(\mathbb{P})$ only (i) when the conditional mean outcome is consistently estimated, which can only occur under the null ($\mathbb{E}(Y \mid A, W^C) = \mathbb{E}(Y \mid W^C)$); *or* (ii) in a randomized trial where the propensity score is known and always consistently estimated. This implies that in an observational setting, CARE is not consistent when there is an intervention effect. However, by controlling for confounders W^C when predicting the outcome, CARE is still expected to be less biased that the unadjusted in observational settings. Second, under the same regularity conditions, CARE is asymptotically normal, and its variance can be estimated by the sample variance of the following influence curve divided by sample size n:

$$\hat{IC}^{CARE} = \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A_i = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A_i = 0)} \right) \left(Y_i - \hat{\mathbb{E}}(Y_i \mid W_i^C) \right) - \hat{\psi}.$$

Finally, under consistent estimation of both nuisance parameters, which can only occur under the null and in a trial setting, CARE will be locally efficient. In trial setting, however, we do expect CARE to provide efficiency gains over the unadjusted estimator from covariate adjustment when predicting the outcome $\hat{\mathbb{E}}(Y \mid W^C)$.

B.3 Supplementary Materials

B.3.1 Diagrams including unmeasured covariates

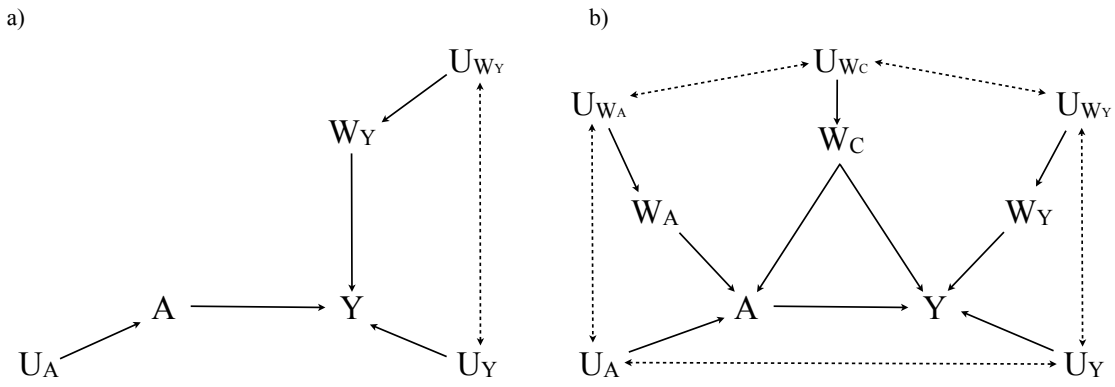

Figure B.1. Causal diagrams for randomized trials **(a)** and observational studies **(b)** including measured and unmeasured covariates. These diagrams give us a visual representation of the relationships between the variables in a causal model. Arrows are drawn from a cause to an effect; dashed double-sided arrows indicate an unknown or unmeasured relationship. In a randomized setting, the exposure of interest (A) is independent of all other variables and the outcome of interest (Y) is influenced by both A and a set of other covariates (W^Y). Randomization also guarantees that the unmeasured factors influencing A (U_A) are independent of the unmeasured factors influencing W^Y (U_{W^Y}) and Y (U_Y). In an observational setting, A is no longer randomized, but instead influenced by other covariates. Some of these covariates (W^C) also influence Y, thus confounding the relationship between A and Y. Other covariates (W^A) only influence A and not Y. Without randomization any of the unmeasured covariates may have a relationship with any of the other unmeasured covariates, as indicated by the dashed arrows around the perimeter of the diagram.

B.3.2 The unadjusted estimator is consistent for the statistical parameter Ψ^{RCT} in randomized trials

Theorem S1: In randomized trials, due to the absence of confounding covariates, the unadjusted estimator is consistent for the statistical parameter Ψ^{RCT}.

Proof: Starting from (3.4), we can use the fact that $n_a = n \times \hat{\mathbb{P}}(A = a)$, where $\hat{\mathbb{P}}(A = a)$ is the empirical likelihood of assignment to exposure a, to rewrite the

estimator like so:

$$\begin{aligned}
\hat{\Psi}^{unadj} &= \frac{1}{n_1} \sum_{i \forall A_i = 1} Y_i - \frac{1}{n_0} \sum_{i \forall A_i = 0} Y_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 1) Y_i}{\hat{\mathbb{P}}(A = 1)} - \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}(A_i = 0) Y_i}{\hat{\mathbb{P}}(A = 0)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{I}(A_i = 1)}{\hat{\mathbb{P}}(A = 1)} - \frac{\mathbb{I}(A_i = 0)}{\hat{\mathbb{P}}(A = 0)} \right) Y_i.
\end{aligned}$$

The expected value of the unadjusted estimator is equal to the statistical parameter Ψ^{RCT} (3.3), which identifies the average treatment effect in randomized trials by design:

$$\begin{aligned}
\mathbb{E}\left[\hat{\Psi}^{unadj}\right] &= \mathbb{E}\left[\frac{\mathbb{I}(A = 1)}{\hat{\mathbb{P}}(A = 1)} Y\right] - \mathbb{E}\left[\frac{\mathbb{I}(A = 0)}{\hat{\mathbb{P}}(A = 0)} Y\right] \\
&= \sum_{w^Y, a, y} \frac{\mathbb{I}(A = 1)}{\hat{\mathbb{P}}(A = 1)} y \mathbb{P}(Y = y \mid A = a, W^Y = w^Y) \mathbb{P}(A = a \mid W^Y = w^Y) \mathbb{P}(W^Y = w^Y) \\
&\quad - \sum_{w^Y, a, y} \frac{\mathbb{I}(A = 0)}{\hat{\mathbb{P}}(A = 0)} y \mathbb{P}(Y = y \mid A = a, W^Y = w^Y) \mathbb{P}(A = a \mid W^Y = w^Y) \mathbb{P}(W^Y = w^Y) \\
&= \sum_{w^Y, y} y \mathbb{P}(Y = y \mid A = 1, W^Y = w^Y) \mathbb{P}(W^Y = w^Y) \\
&\quad - \sum_{w^Y, y} y \mathbb{P}(Y = y \mid A = 0, W^Y = w^Y) \mathbb{P}(W^Y = w^Y) \\
&= \mathbb{E}_{W^Y}\Big[\mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0)\Big] \\
&= \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) \\
&= \Psi^{RCT},
\end{aligned}$$

where summations generalize to integrals for continuous random variables. By satisfying the randomization assumption, $\mathbb{P}(A = a \mid W^Y = W^Y) = \mathbb{P}(A = a)$ which cancels with the empirical probability of allocation $\hat{\mathbb{P}}(A = a)$ for each group. Since the baseline covariates W^Y are independent of the exposure A, their distribution is asymptotically equivalent between exposure groups.

B.3.3 The unadjusted estimator is biased for the statistical parameter Ψ in observational studies

Theorem S2: In observational settings, when there are covariates that influence both the allocation to exposure and the outcome, the unadjusted estimator is confounded and is thus biased for the statistical parameter Ψ.

Proof:

$$\begin{aligned}
\mathbb{E}\left[\hat{\Psi}^{unadj}\right] &= \mathbb{E}\left[\frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1)}Y\right] - \mathbb{E}\left[\frac{\mathbb{I}(A=0)}{\hat{\mathbb{P}}(A=0)}Y\right] \\
&= \sum_{w,a,y}\frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1)}y\mathbb{P}(Y=y \mid A=a, W=w)\mathbb{P}(A=a \mid W=w)\mathbb{P}(W=w) \\
&\quad - \sum_{w,a,y}\frac{\mathbb{I}(A=0)}{\hat{\mathbb{P}}(A=0)}y\mathbb{P}(Y=y \mid A=a, W=w)\mathbb{P}(A=a \mid W=w)\mathbb{P}(W=w) \\
&= \sum_{w,y}\frac{\mathbb{P}(A=1 \mid W^C=w^C)}{\hat{\mathbb{P}}(A=1)}y\mathbb{P}(Y=y \mid A=1, W=w)\mathbb{P}(W=w) \\
&\quad - \sum_{w,y}\frac{\mathbb{P}(A=0 \mid W^C=w^C)}{\hat{\mathbb{P}}(A=0)}y\mathbb{P}(Y=y \mid A=0, W=w)\mathbb{P}(W=w) \\
&\neq \Psi,
\end{aligned}$$

where the set of baseline covariates W contains both the covariates that only influence the outcome W^Y and the confounding covariates W^C. In this setting, the conditional probability of exposure depends on the confounding covariates and is independent of the other baseline covariates W^Y that only affect the outcome $\mathbb{P}(A=a \mid W=w) = \mathbb{P}(A=a \mid W^C=W^C)$. Due to this confounding, the conditional probability of exposure is not offset by the empirical probability of exposure $\hat{\mathbb{P}}(A=1)$ as in randomized settings.

B.3.4 The IPTW estimator is consistent for the statistical parameter Ψ

Theorem S3: In randomized or observational settings, the inverse probability weighting estimator is consistent for the statistical parameter Ψ when propensity scores are consistently estimated.

Proof: The inverse probability of treatment weighting (IPTW) estimator adjusts for the confounding covariates between the outcome and the allocation to exposure by replacing the empirical probability of exposure $\hat{\mathbb{P}}(A = 1)$ with propensity scores $\hat{\mathbb{P}}(A = 1 \mid W^C = w^C)$. If the propensity scores are consistent for the true conditional probability of exposure $\mathbb{P}(A = 1 \mid W^C = w^C)$, the IPTW estimator makes consistent estimates for the statistical parameter Ψ:

$$
\begin{aligned}
\mathbb{E}\left[\hat{\Psi}^{IPTW}\right] &= \mathbb{E}\left[\frac{\mathbb{I}(A=1)}{\hat{\mathbb{P}}(A=1 \mid W^C = w^C)}Y\right] - \mathbb{E}\left[\frac{\mathbb{I}(A=0)}{1-\hat{\mathbb{P}}(A=1 \mid W^C = w^C)}Y\right] \\
&= \sum_{w,a,y} \frac{\mathbb{I}(A=1)y\mathbb{P}(Y=y \mid A=a, W=w)\mathbb{P}(A=a \mid W=w)\mathbb{P}(W=w)}{\hat{\mathbb{P}}(A=1 \mid W^C = w^C)} \\
&\quad - \sum_{w,a,y} \frac{\mathbb{I}(A=0)y\mathbb{P}(Y=y \mid A=a, W=w)\mathbb{P}(A=a \mid W=w)\mathbb{P}(W=w)}{1-\hat{\mathbb{P}}(A=1 \mid W^C = w^C)} \\
&= \sum_{w,a,y} \frac{\mathbb{I}(A=1)y\mathbb{P}(Y=y \mid A=a, W=w)\mathbb{P}(A=a \mid W^C=w^C)\mathbb{P}(W=w)}{\hat{\mathbb{P}}(A=1 \mid W^C = w^C)} \\
&\quad - \sum_{w,a,y} \frac{\mathbb{I}(A=0)y\mathbb{P}(Y=y \mid A=a, W=w)\mathbb{P}(A=a \mid W^C=w^C)\mathbb{P}(W=w)}{1-\hat{\mathbb{P}}(A=1 \mid W^C = w^C)} \\
&= \sum_{w,y} y\mathbb{P}(Y=y \mid A=1, W=w)\mathbb{P}(W=w) - \sum_{w,y} y\mathbb{P}(Y=y \mid A=0, W=w)\mathbb{P}(W=w) \\
&= \mathbb{E}_W\Big[\mathbb{E}(Y \mid A=1) - \mathbb{E}(Y \mid A=0)\Big] \\
&= \mathbb{E}_{W^C}\Big[\mathbb{E}(Y \mid A=1) - \mathbb{E}(Y \mid A=0)\Big] \\
&= \Psi.
\end{aligned}
$$

If there is no unmeasured confounding and the propensity scores are consistent for the true probability of allocation, then they cancel out each other out. The distribution of

covariates that only influence the outcome W^Y are asymptotically equivalent between the exposure groups, therefore the expectation only needs to be taken over the confounding covariates W^C. The IPTW estimator is especially sensitive to positivity violations. The propensity scores must be between zero and one across all values of $W^C = w^C$, otherwise $\hat{\Psi}^{IPTW}$ will be undefined. Note that the propensity score is not required to include the baseline covariates that influence only the allocation to exposure W^A for this estimator to be consistent for Ψ. In fact, including W^A in the estimation of the propensity scores could make predictions for A that are closer to zero or one, which could destabilize the estimate of the exposure effect.

The statistical parameter Ψ identifies the ATE in randomized and observational settings under the identifiability assumptions outlined in Section 3.2.

B.3.5 More simulation results

B.3.5.1 Ghana sims

In randomized trials with the estimators restricted to using only average age in months $W1$ and percent of children who are female $W2$, as in the case study, all estimators had low bias and sufficiently high confidence interval coverage. but CARE and CARE–IPW had more statistical power and less variance (Table B.1). These simulations suggest that the CARE and CARE–IPW estimators may add value in the case study due to their more precise estimates of the exposure effect.

B.3.5.2 Full observational results

When not accounting for the confounding covariate prior childhood mortality $W3$, the IPTW estimator had large bias in the observational simulations (Table B.2). When accounting for the confounding covariate, the IPTW estimator had slightly more bias but less variability when accounting for non-confounding covariates average age in months $W1$ and percent of children that were female $W2$. The IPTW estimator had high confidence interval coverage regardless of the covariates it adjusted for.

Exposure	Estimator	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
Effect	CARE–IPW	0.03	4.85	4.71	94.2%	69.6%
	CARE	0.28	4.73	4.65	94.4%	69%
	IPTW	0.02	4.84	8.91	100%	12.3%
	Unadj	0.01	5.35	5.24	94.4%	60.6%
Null	CARE–IPW	0.02	5.01	4.86	94.3%	5.7%
	CARE	0.03	4.89	4.80	94.5%	5.5%
	IPTW	0.02	5.01	10.01	100%	0%
	Unadj	0.01	5.53	5.42	94.4%	5.6%

Table B.1. Simulation results for the effect estimators in randomized trials, with the regressions for predicting the outcome and estimating the propensity scores restricted to using $W1$ and $W2$ as in the bednets case study.

Exposure	P-score	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
Effect	W1+W2	8.63	4.77	8.86	96.7%	0.2%
	W1+W2+W3	0.36	5.70	10.23	99.9%	2.7%
	W3	0.26	6.10	10.08	99.9%	5.1%
Null	W1+W2	9.35	4.93	10.04	98.3%	1.7%
	W1+W2+W3	0.37	5.80	11.21	100%	0%
	W3	0.27	6.23	11.05	100%	0%

Table B.2. Simulation results for the IPTW estimator in observational settings across approaches.

When CARE did not account for the confounding covariate $W3$, it was biased with low confidence interval coverage (Table B.3). When there was an exposure effect, CARE had less bias and more statistical power when it accounted for all covariates than when it only accounted for the confounding covariate. However, under the null, CARE had more bias when it accounted for all covariates than when it only accounted for the confounding covariate.

When CARE–IPW did not account for the confounding covariate $W3$ in its propensity score, it had similar results to CARE for each outcome equation (Table B.4). When CARE–IPW accounted for the confounding covariate in the propensity

Exposure	Outcome	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
Effect	W1+W2	8.69	4.68	4.59	51.7%	11.6%
	W1+W2+W3	0.10	4.08	4.20	95.2%	80.6%
	W3	1.45	4.36	4.56	94.9%	61.9%
Null	W1+W2	9.14	4.83	4.75	50.6%	49.4%
	W1+W2+W3	-1.51	4.04	4.25	95.1%	4.9%
	W3	0.13	4.40	4.62	95.8%	4.2%

Table B.3. Simulation results for CARE in observational settings across approaches.

score, with or without the non-confounding covariates, the estimator had low bias and high confidence interval coverage. When accounting for all of the covariates to predict the outcome, not including the exposure, the CARE–IPW estimator had the least average standard error and the most statistical power.

Exposure	Outcome	P-score	Bias	MC SE	Average SE	CI coverage	Power/ Type I error
	W1+W2	W1+W2	8.62	4.79	4.65	53%	12.1%
	W1+W2+W3	W1+W2	-0.19	4.19	4.28	95.3%	81.2%
	W3	W1+W2	1.48	3.87	4.63	97%	62.3%
	W1+W2	W1+W2+W3	0.23	4.82	5.68	97.2%	54.2%
Effect	W1+W2+W3	W1+W2+W3	-0.16	5.02	4.64	94.9%	75.4%
	W3	W1+W2+W3	0.05	4.50	5.06	97.5%	67.8%
	W1+W2	W3	0.44	4.75	5.57	97.1%	54.5%
	W1+W2+W3	W3	0.16	4.93	4.57	94.9%	74.6%
	W3	W3	-0.02	5.08	4.98	94.9%	67.3%
	W1+W2	W1+W2	9.33	4.95	4.81	49.6%	50.4%
	W1+W2+W3	W1+W2	-1.52	4.14	4.31	95%	5%
	W3	W1+W2	0.16	3.87	4.70	98%	2%
	W1+W2	W1+W2+W3	0.21	4.78	5.52	97.2%	2.8%
Null	W1+W2+W3	W1+W2+W3	-0.16	5.58	4.74	94.4%	5.6%
	W3	W1+W2+W3	-0.05	4.69	5.12	97.5%	2.5%
	W1+W2	W3	0.18	4.72	5.45	97.2%	2.8%
	W1+W2+W3	W3	-0.15	5.28	4.68	94.4%	5.6%
	W3	W3	-0.11	5.28	5.03	94.6%	5.4%

Table B.4. Simulation results for CARE–IPW in observational settings across approaches.

APPENDIX C

INCORPORATING FORECASTS INTO ESTIMATES OF THE AVERAGE TREATMENT EFFECT WITH AN APPLICATION TO ZIKA EMERGENCY OPERATIONS CENTERS IN THAILAND SUPPLEMENTAL MATERIALS

C.1 Synthetic simulation correlation matrix Σ

		$Temp$			$Rain$			DHF		
		N	D	J	N	D	J	N	D	J
$Temp$	N	1	0.8	0.6	0.4	0.2	0.1	0.4	0.2	0.1
	D	0.8	1	0.8	0.2	0.4	0.2	0.2	0.4	0.2
	J	0.6	0.8	1	0.1	0.2	0.4	0.1	0.2	0.4
$Rain$	N	0.4	0.2	0.1	1	0.8	0.6	0.4	0.2	0.1
	D	0.2	0.4	0.2	0.8	1	0.8	0.2	0.4	0.2
	J	0.1	0.2	0.4	0.6	0.8	1	0.1	0.2	0.4
DHF	N	0.4	0.2	0.1	0.4	0.2	0.1	1	0.8	0.6
	D	0.2	0.4	0.2	0.2	0.4	0.2	0.8	1	0.8
	J	0.1	0.2	0.4	0.1	0.2	0.4	0.6	0.8	1

Table C.1. The correlation matrix Σ used to generate synthetic values of temperature, rainfall, and dengue hemorrhagic fever (DHF) incidence for November, December, and January (denoted N, D, and J).

C.2 Forecasting accuracy and effect estimation

C.2.1 Synthetic simulation

C.2.1.1 IPTW

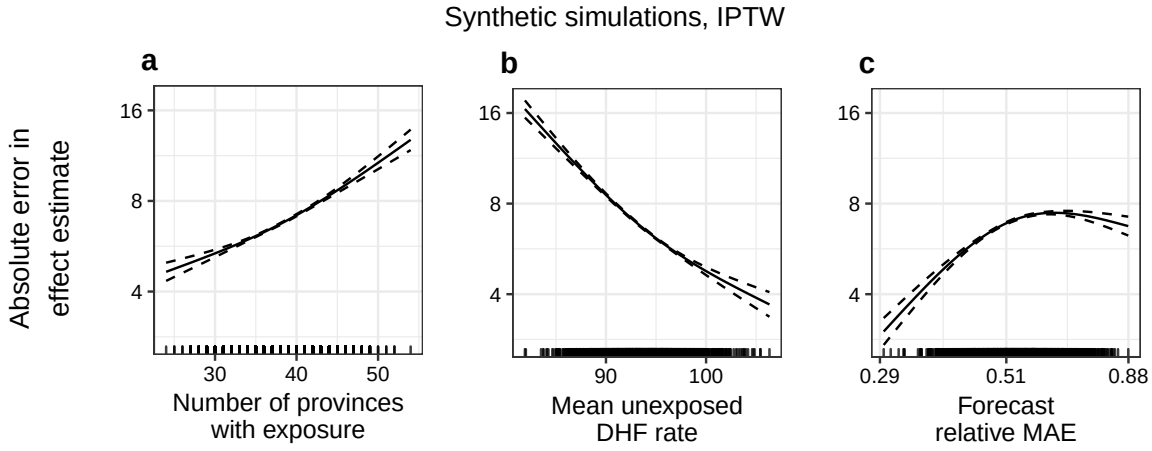

Figure C.1. The covariate fit curves showing the associations between the absolute error in effect estimation with the IPTW(F) estimator and simulation-level covariates including forecasting relative mean absolute error in the synthetic simulations. Simulations with fewer provinces receiving the exposure, higher DHF incidence amongst unexposed provinces, and with less forecasting error relative to the baseline had less estimation error.

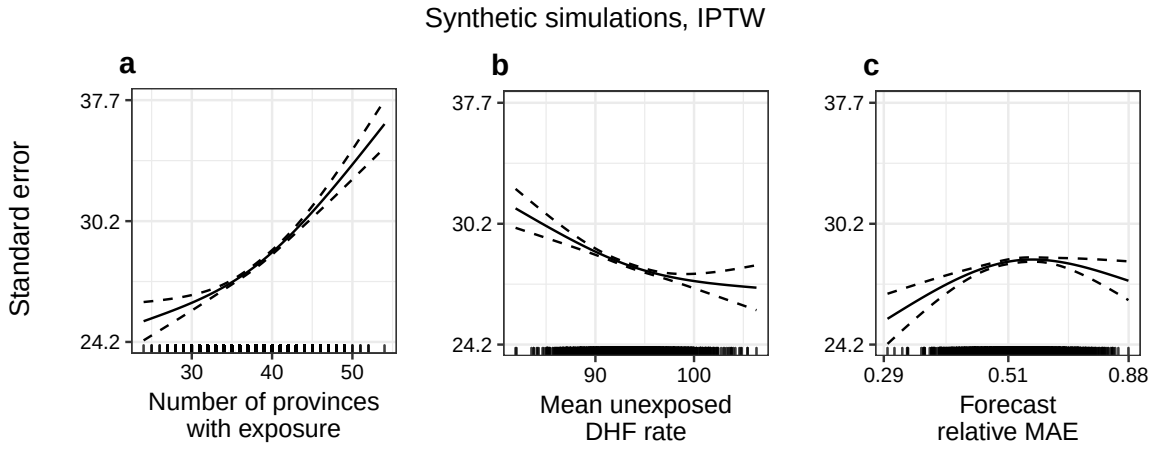

Figure C.2. The covariate fit curves showing the associations between the standard error of the IPTW(F) estimator and simulation-level covariates including forecasting relative mean absolute error in the synthetic simulations. Simulations with fewer provinces receiving the exposure, higher DHF incidence amongst unexposed provinces, and with less forecasting error relative to the baseline had less estimated standard error.

C.2.1.2 CARE–IPW

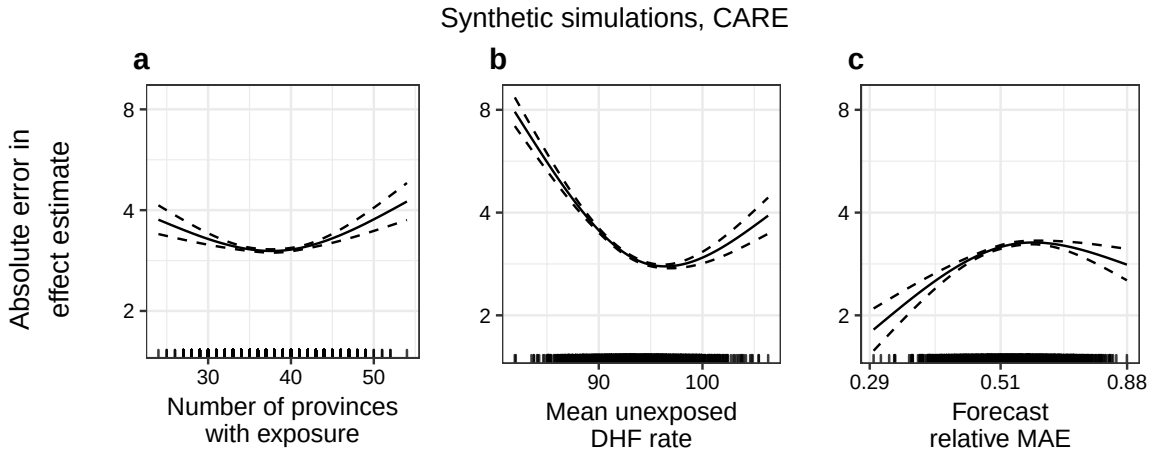

Figure C.3. The covariate fit curves showing the associations between the absolute error in effect estimation with the CARE(F,F) estimator and simulation-level covariates including forecasting relative mean absolute error in the synthetic simulations. After adjusting for the number of exposed provinces and the unexposed DHF rate, less forecast relative mean absolute error reduced the error in effect estimation.

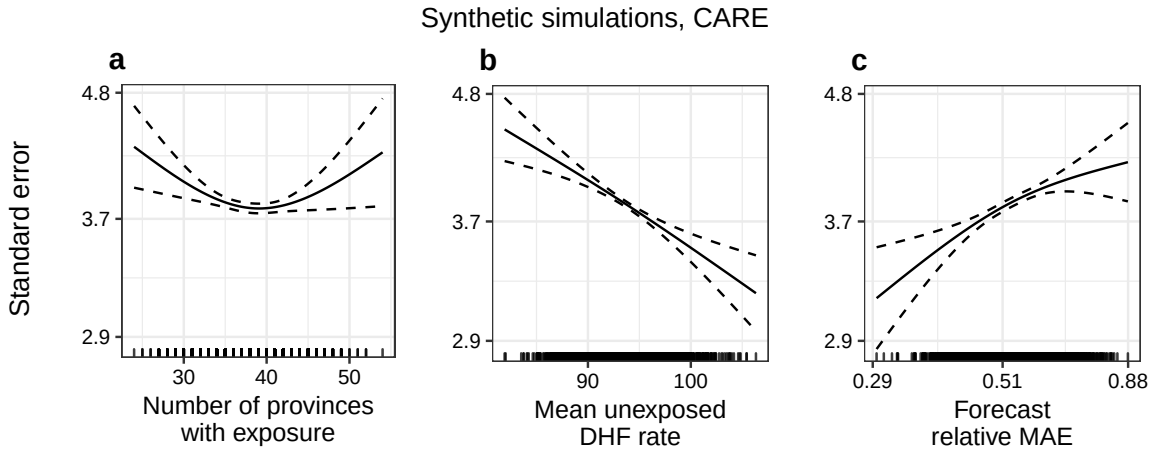

Figure C.4. The covariate fit curves showing the associations between the standard error of the CARE(F,F) estimator and simulation-level covariates including forecasting relative mean absolute error in the synthetic simulations. Simulations with higher DHF incidence amongst unexposed provinces and with less forecasting error relative to the baseline had less estimated standard error.

C.2.2 Historical simulation

C.2.2.1 IPTW

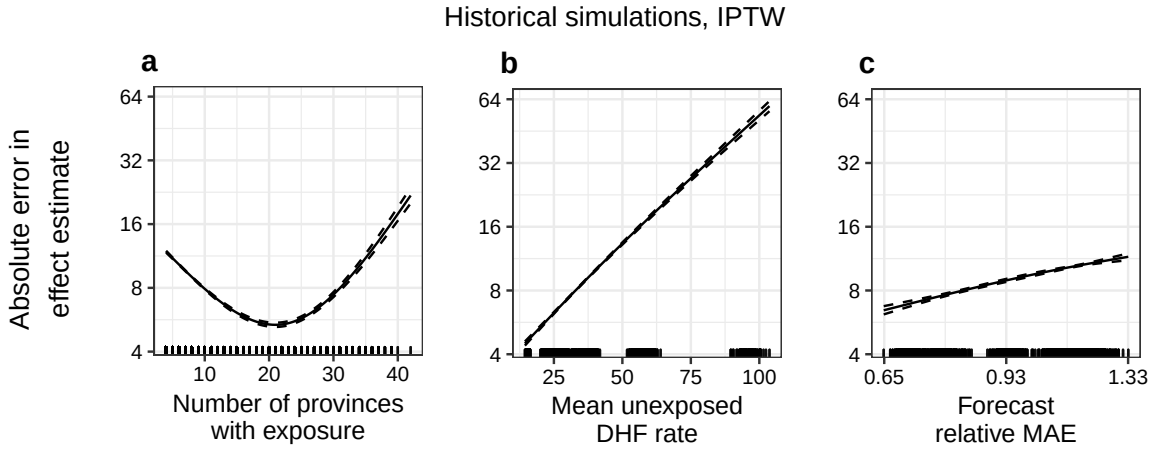

Figure C.5. The covariate fit curves showing the associations between the absolute error in effect estimation with the IPTW(F) estimator and simulation-level covariates including forecasting relative mean absolute error in the historical simulations. Simulations with lower DHF incidence amongst unexposed provinces and with less forecasting error relative to the baseline had less error in effect estimates.

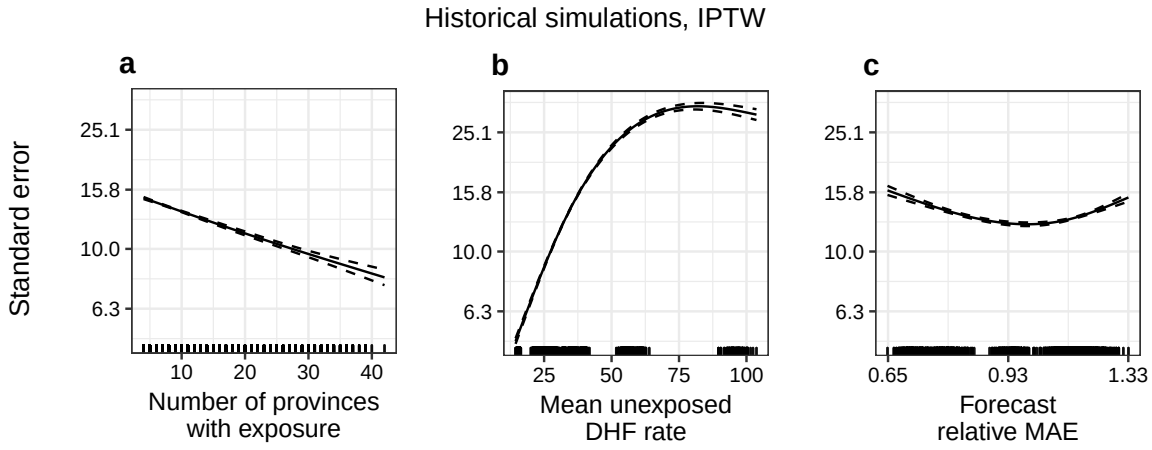

Figure C.6. The covariate fit curves showing the associations between the standard error of the IPTW(F) estimator and simulation-level covariates including forecasting relative mean absolute error in the historical simulations. After adjusting for the number of exposed provinces and the unexposed DHF rate, forecast relative mean absolute error had a negligible association with estimated standard error.

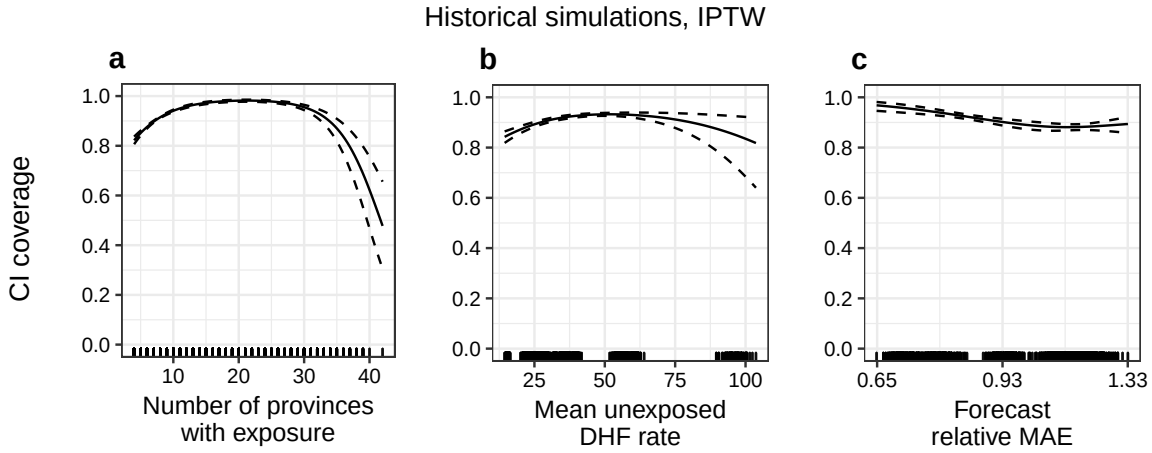

Figure C.7. The covariate fit curves showing the associations between the confidence interval coverage of the IPTW(F) estimator and simulation-level covariates including forecasting relative mean absolute error in the historical simulations. After adjusting for the number of exposed provinces and the unexposed DHF rate, less forecast relative mean absolute error was associated with higher confidence interval coverage.

C.2.2.2 CARE–IPW

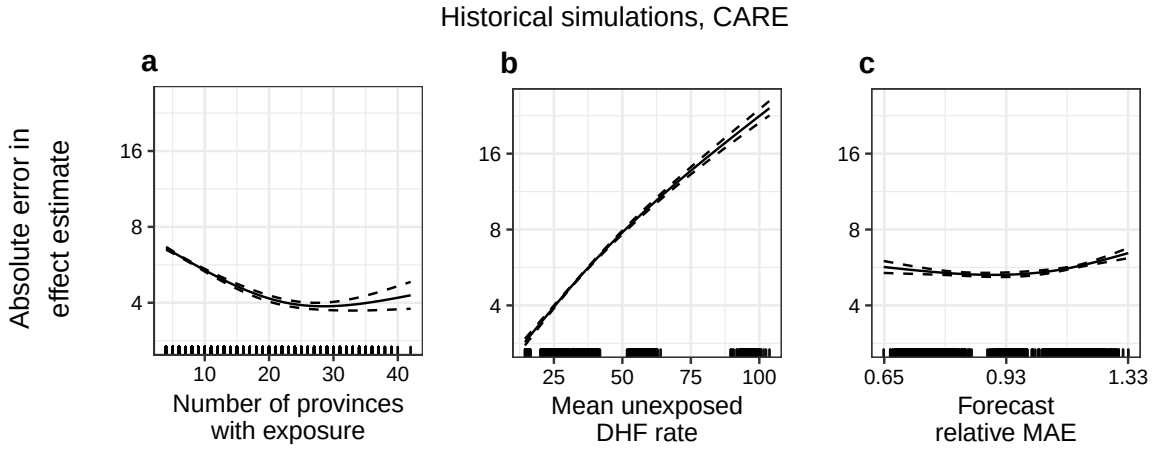

Figure C.8. The covariate fit curves showing the associations between the absolute error in effect estimation with the CARE(L,F) estimator with respect to simulation-level covariates including forecasting relative mean absolute error in the historical simulations. After adjusting for the number of exposed provinces and the unexposed DHF rate, less forecast relative mean absolute error was associated slightly less error in effect estimation.

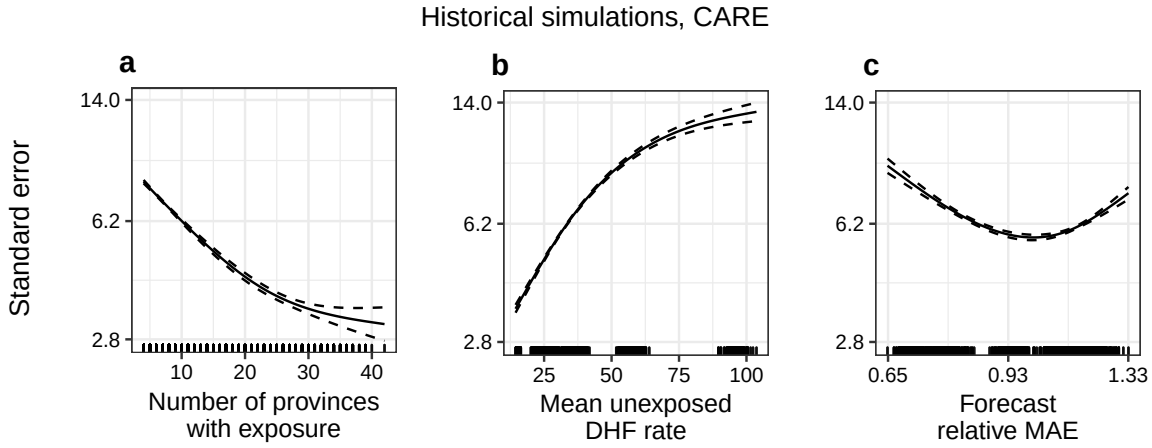

Figure C.9. The covariate fit curves showing the associations between the standard error of the CARE(L,F) estimator and simulation-level covariates including forecasting relative mean absolute error in the synthetic simulations.

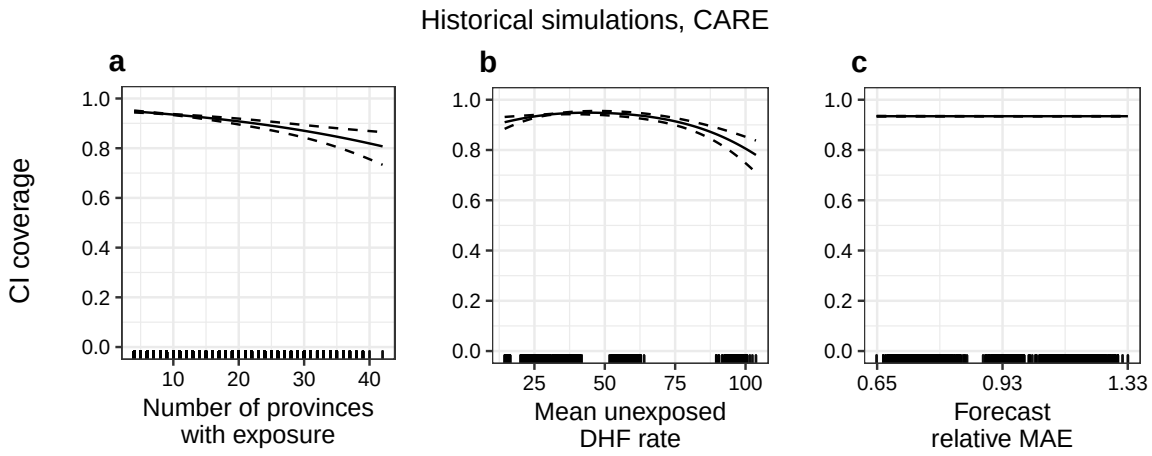

Figure C.10. The covariate fit curves showing the associations between the confidence interval coverage of the CARE(L,F) estimator and simulation-level covariates including forecasting relative mean absolute error in the historical simulations. Confidence interval coverage had no association with forecasting error after adjusting for the number of exposed provinces and the unexposed DHF rate.

BIBLIOGRAPHY

[1] Situation of zika virus disease in thailand – 29 june 2559. Tech. rep., Bureau of Emerging Infectious Diseases, 2016.

[2] Zika virus control handbook for healthcare providers and public health officials 2016. Tech. rep., Bureau of Emerging Infectious Diseases, Department of Disease control, Ministry of Public Health, Thailand, Nonthaburi, Thailand, 2016.

[3] Zika situation report. Tech. rep., World Health Organization, 2017.

[4] Adams, B., Holmes, E. C., Zhang, C., Mammen, M. P., Nimmannitya, S., Kalayanarooj, S., and Boots, M. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in bangkok. *Proceedings of the National Academy of Sciences 103*, 38 (Sept. 2006), 14234–14239.

[5] Adler, Robert F., Huffman, George J., Chang, Alfred, Ferraro, Ralph, Xie, Ping-Ping, Janowiak, John, Rudolf, Bruno, Schneider, Udo, Curtis, Scott, Bolvin, David, Gruber, Arnold, Susskind, Joel, Arkin, Philip, and Nelkin, Eric. The version-2 global precipitation climatology project (gpcp) monthly precipitation analysis (1979–present). *Journal of Hydrometeorology 4*, 6 (Dec. 2003), 1147–1167.

[6] Armstrong, J Scott. Evaluating forecasting methods. *International Journal of Forecasting*, 1990 (2001), 443–472.

[7] Asher, Jason, Barker, Christopher, Chen, Grace, Cummings, Derek, Chinazzi, Matteo, Daniel-Wayman, Shelby, Fischer, Marc, Ferguson, Neil, Follman, Dean, Halloran, M. Elizabeth, Johansson, Michael, Kugeler, Kiersten, Kwan, Jennifer, Lessler, Justin, Longini, Ira M., Merler, Stefano, Monaghan, Andrew, Piontti, Ana Pastore y, Perkins, Alex, Prevots, D. Rebecca, Reiner, Robert, Rossi, Luca, Rodriguez-Barraquer, Isabel, Siraj, Amir S., Sun, Kaiyuan, Vespignani, Alessandro, and Zhang, Qian. Preliminary results of models to predict areas in the americas with increased likelihood of zika virus transmission in 2017. *bioRxiv* (Sept. 2017), 187591.

[8] Babyak, Michael A. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine 66*, 3 (2004), 411–421.

[9] Balzer, Laura B, van der Laan, Mark J, Petersen, Maya L, and SEARCH Collaboration, the SEARCH. Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in medicine 35*, 25 (2016), 4528–4545.

[10] Bennett, Steve, Parpia, Tamiza, Hayes, Richard, and Cousens, Simon. Methods for the analysis of incidence rates in cluster randomized trials. *International Journal of Epidemiology 31*, 4 (Aug. 2002), 839–846.

[11] Bergmeir, Christoph, and Benítez, José M. On the use of cross-validation for time series predictor evaluation. *Information Sciences 191* (May 2012), 192–213.

[12] Bergmeir, Christoph, Hyndman, Rob J., and Koo, Bonsoo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis 120* (Apr. 2018), 70–83.

[13] Bhatt, Samir, Gething, Peter W., Brady, Oliver J., Messina, Jane P., Farlow, Andrew W., Moyes, Catherine L., Drake, John M., Brownstein, John S., Hoen, Anne G., Sankoh, Osman, Myers, Monica F., George, Dylan B., Jaenisch, Thomas, and Wint, G. R. William. The global distribution and burden of dengue. *Nature 496*, 7446 (2013), 504–507.

[14] Bickel, Peter J, Klaassen, C AJ, Ritov, Y, and Wellner, J A. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, 1993.

[15] Biggerstaff, Matthew, Alper, David, Dredze, Mark, Fox, Spencer, Fung, Isaac Chun-Hai, Hickmann, Kyle S., Lewis, Bryan, Rosenfeld, Roni, Shaman, Jeffrey, Tsou, Ming-Hsiang, Velardi, Paola, Vespignani, Alessandro, and Finelli, Lyn. Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge. *BMC Infectious Diseases 16*, 1 (Dec. 2016), 357.

[16] Binka, F. N., Kubaje, A., Adjuik, M., Williams, L. A., Lengeler, C., Maude, G. H., Armah, G. E., Kajihara, B., Adiamah, J. H., and Smith, P. G. Impact of permethrin impregnated bednets on child mortality in kassena-nankana district, ghana: a randomized controlled trial. *Tropical Medicine & International Health 1*, 2 (1996), 147–154.

[17] Birrell, P J, Ketsetzis, G, Gay, N J, Cooper, B S, Presanis, A M, Harris, R J, Charlett, A, Zhang, X S, White, P J, Pebody, R G, and De Angelis, D. Bayesian modeling to unmask and predict influenza a/h1n1pdm dynamics in london. *Proc Natl Acad Sci U S A 108*, 45 (2011), 18238–18243.

[18] Black, N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ (Clinical research ed.) 312*, 7040 (May 1996), 1215–8.

[19] Blackwell, G. L. A potential multivariate index of condition for small mammals. *New Zealand Journal of Zoology 29*, 3 (Jan. 2002), 195–203.

[20] Bogner, Konrad, Liechti, Katharina, Bernhard, Luzi, Monhart, Samuel, and Zappa, Massimiliano. Skill of hydrological extended range forecasts for water resources management in switzerland. *Water Resources Management 32*, 3 (Feb. 2018), 969–984.

[21] Bonačić Marinović, Axel, Swaan, Corien, van Steenbergen, Jim, and Kretzschmar, Mirjam. Quantifying reporting timeliness to improve outbreak control. *Emerging infectious diseases 21*, 2 (Feb. 2015), 209–16.

[22] Box, GEP, and Jenkins, GM. Some statistical aspects of adaptive optimization and control. *Journal of the Royal Statistical Society. Series B (... 24*, 2 (1962), 297–343.

[23] Bradley, A. Allen, and Schwartz, Stuart S. Summary verification measures and their interpretation for ensemble forecasts. *Monthly Weather Review 139*, 9 (Sept. 2011), 3075–3089.

[24] Brady, Oliver J., Johansson, Michael A., Guerra, Carlos A., Bhatt, Samir, Golding, Nick, Pigott, David M., Delatte, Hélène, Grech, Marta G., Leisnham, Paul T., Maciel-de Freitas, Rafael, Styer, Linda M., Smith, David L., Scott, Thomas W., Gething, Peter W., and Hay, Simon I. Modelling adult aedes aegypti and aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & Vectors 6*, 1 (2013), 351.

[25] Bray, Andrew, and Schoenberg, Frederic Paik. Assessment of point process models for earthquake forecasting. *Statistical Science 28*, 4 (Nov. 2013), 510–520.

[26] Buczak, Anna L., Baugher, Benjamin, Moniz, Linda J., Bagley, Thomas, Babin, Steven M., and Guven, Erhan. Ensemble method for dengue prediction. *PLOS ONE 13*, 1 (Jan. 2018), e0189988.

[27] Buczak, Anna L, Koshute, Phillip T, Babin, Steven M, Feighner, Brian H, and Lewis, Sheryl H. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making 12* (2012), 124.

[28] Burke, Donald S., Nisalak, Ananda, Johnson, David E., and Scott, Robert McN. A prospective study of dengue infections in bangkok. *American Journal of Tropical Medicine and Hygiene 38*, 1 (Jan. 1988), 172–180.

[29] Campbell, James E. Polls and votes. *American Politics Quarterly 24*, 4 (Oct. 1996), 408–433.

[30] Campbell, Karen M., Lin, C. D., Iamsirithaworn, Sopon, and Scott, Thomas W. The complex relationship between weather and dengue virus transmission in thailand. *American Journal of Tropical Medicine and Hygiene 89*, 6 (2013), 1066–1080.

[31] Cauchemez, Simon, Besnard, Marianne, Bompard, Priscillia, Dub, Timothée, Guillemette-Artur, Prisca, Eyrolle-Guignot, Dominique, Salje, Henrik, Van Kerkhove, Maria D, Abadie, Véronique, Garel, Catherine, Fontanet, Arnaud, and Mallet, Henri-Pierre. Association between zika virus and microcephaly in french polynesia, 2013–15: a retrospective study. *The Lancet 387*, 10033 (May 2016), 2125–2132.

[32] Climatic, National, Squadron, Weather, Meteorology, Fleet Numerical, and Detachment, Oceanography. Federal climate complex data documentation for integrated surface data. Tech. rep., National Climatic Data Center, Asheville, NC, 2015.

[33] Cochran, William G. Analysis of covariance: Its nature and uses. *Biometrics 13*, 3 (Sept. 1957), 261.

[34] Cox, D. R., and McCullagh, P. A biometrics invited paper with discussion. some aspects of analysis of covariance. *Biometrics 38*, 3 (Sept. 1982), 541.

[35] Danielson-François, Anne, Fetterer, Christine A., and Smallwood, Peter D. Body condition and mate choice in tetragnatha elongata (araneae, tetragnathidae). *http://dx.doi.org/10.1636/0161-8202(2002)030[0020:BCAMCI]2.0.CO;2* (Jan. 2009).

[36] DARPA. Chikv challenge announces winners, progress toward forecasting the spread of infectious diseases, 2015.

[37] Deiner, Michael S., Worden, Lee, Rittel, Alex, Ackley, Sarah F., Liu, Fengchen, Blum, Laura, Scott, James C., Lietman, Thomas M., and Porco, Travis C. Short-term leprosy forecasting from an expert opinion survey. *PLOS ONE 12*, 8 (Aug. 2017), e0182245.

[38] Dejnirattisai, Wanwisa, Supasa, Piyada, Wongwiwat, Wiyada, Rouvinski, Alexander, Barba-Spaeth, Giovanna, Duangchinda, Thaneeya, Sakuntabhai, Anavaj, Cao-Lormeau, Van-Mai, Malasit, Prida, Rey, Felix A, Mongkolsapaya, Juthathip, and Screaton, Gavin R. Dengue virus sero-cross-reactivity drives antibody-dependent enhancement of infection with zika virus. *Nature Immunology 17*, 9 (Sept. 2016), 1102–1108.

[39] DeRoeck, Denise, Deen, Jacqueline, and Clemens, John D. Policymakers' views on dengue fever/dengue haemorrhagic fever and the need for dengue vaccines in four southeast asian countries. *Vaccine 22*, 1 (Dec. 2003), 121–129.

[40] Diebold, Francis X. *Elements of forecasting*, 4th ed. Department of Economics, University of Pennsylvania, 2001.

[41] dos Santos, Thais, Rodriguez, Angel, Almiron, Maria, Sanhueza, Antonio, Ramon, Pilar, de Oliveira, Wanderson K., Coelho, Giovanini E., Badaró, Roberto, Cortez, Juan, Ospina, Martha, Pimentel, Raquel, Masis, Rolando, Hernandez, Franklin, Lara, Bredy, Montoya, Romeo, Jubithana, Beatrix, Melchor, Angel, Alvarez, Angel, Aldighieri, Sylvain, Dye, Christopher, and Espinal, Marcos A. Zika virus and the guillain–barré syndrome — case series from seven countries. *New England Journal of Medicine 375*, 16 (Oct. 2016), 1598–1601.

[42] Draper, Norman R., and Smith, Harry. *Applied Regression Analysis*. Wiley, New York, 1998.

[43] Du, Xiangjun, King, Aaron A, Woods, Robert J, and Pascual, Mercedes. Evolution-informed forecasting of seasonal influenza a (h3n2). *Science translational medicine 9*, 413 (Oct. 2017), eaan5325.

[44] Du, Xiangjun, and Pascual, Mercedes. Incidence prediction for the 2017-2018 influenza season in the united states with an evolution-informed model. *PLoS Currents* (2018).

[45] Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., and Rothman, R. E. Google flu trends: Correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases 54*, 4 (Feb. 2012), 463–469.

[46] Duncan, Otis Dudley. *Introduction to Structural Equation Models*. Academic Press, Inc., New York, 1975.

[47] Elkin, C. M., and Reid, M. L. Low energy reserves and energy allocation decisions affect reproduction by mountain pine beetles, dendroctonus ponderosae. *Functional Ecology 19*, 1 (Feb. 2005), 102–109.

[48] Endy, Timothy P., Chunsuttiwat, Supamit, Nisalak, Ananda, Libraty, Daniel H., Green, Sharone, Rothman, Alan L., Vaughn, David W., and Ennis, Francis A. Epidemiology of inapparent and symptomatic acute dengue virus infection: A prospective study of primary school children in kamphaeng phet, thailand. *American Journal of Epidemiology 156*, 1 (July 2002), 40–51.

[49] Eubank, S, Guclu, H, Kumar, V S, Marathe, M V, Srinivasan, A, Toroczkai, Z, and Wang, N. Modelling disease outbreaks in realistic urban social networks. *Nature 429*, 6988 (May 2004), 180–184.

[50] Fan, Yun, and van den Dool, Huug. A global monthly land surface air temperature analysis for 1948-present. *Journal of Geophysical Research Atmospheres 113*, 1 (2008).

[51] Farrow, David C., Brooks, Logan C., Hyun, Sangwon, Tibshirani, Ryan J., Burke, Donald S., and Rosenfeld, Roni. A human judgment approach to epidemiological forecasting. *PLOS Computational Biology 13*, 3 (Mar. 2017), e1005248.

[52] Ferguson, Neil M, Cummings, Derek A T, Fraser, Christophe, Cajka, James C, Cooley, Philip C, and Burke, Donald S. Strategies for mitigating an influenza pandemic. *Nature 442*, 7101 (July 2006), 448–452.

[53] Ferguson, Neil M., Rodríguez-Barraquer, Isabel, Dorigatti, Ilaria, Mier-y Teran-Romero, Luis, Laydon, Daniel J., and Cummings, Derek A. T. Benefits and risks of the sanofi-pasteur dengue vaccine: Modeling optimal deployment. *Science 353*, 6303 (Sept. 2016), 1033–1036.

[54] Field, E. H., Dawson, T. E., Felzer, K. R., Frankel, A. D., Gupta, V., Jordan, T. H., Parsons, T., Petersen, M. D., Stein, R. S., Weldon, R. J., and Wills, C. J. Uniform california earthquake rupture forecast, version 2 (ucerf 2). *Bulletin of the Seismological Society of America 99*, 4 (Aug. 2009), 2053–2107.

[55] Fisher, R A. *Statistical methods for research workers*, 4th ed. Oliver and Boyd, Edinburgh, 1932.

[56] Forshey, Brett M., Reiner, Robert C., Olkowski, Sandra, Morrison, Amy C., Espinoza, Angelica, Long, Kanya C., Vilcarromero, Stalin, Casanova, Wilma, Wearing, Helen J., Halsey, Eric S., Kochel, Tadeusz J., Scott, Thomas W., and Stoddard, Steven T. Incomplete protection against dengue virus type 2 re-infection in peru. *PLOS Neglected Tropical Diseases 10*, 2 (Feb. 2016), e0004398.

[57] Funk, Sebastian, Camacho, Anton, Kucharski, Adam J., Eggo, Rosalind M., and Edmunds, W. John. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics 22* (Mar. 2018), 56–61.

[58] Funk, Sebastian, Camacho, Anton, Kucharski, Adam J., Lowe, Rachel, Eggo, Rosalind M., and Edmunds, W. John. Assessing the performance of real-time epidemic forecasts. *bioRxiv* (Aug. 2017), 177451.

[59] Gaffey, Robert H., and Viboud, Cécile. Application of the cdc ebolaresponse modeling tool to disease predictions. *Epidemics 22* (Mar. 2018), 22–28.

[60] Gail, Mitchell H., Mark, Steven D., Carroll, Raymond J., Green, Sylvan B., and Pee, David. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine 15*, 11 (June 1996), 1069–1092.

[61] García-Berthou, Emili. On the misuse of residuals in ecology: testing regression residuals vs. the analysis of covariance. *Journal of Animal Ecology 70*, 4 (July 2001), 708–711.

[62] Gelper, Sarah, Fried, Roland, and Croux, Christophe. Robust forecasting with exponential and holt-winters smoothing. *Journal of Forecasting 29*, 3 (Apr. 2009), 285–300.

[63] Gerardi, D. O., and Monteiro, L. H. A. System identification and prediction of dengue fever incidence in rio de janeiro. *Mathematical Problems in Engineering 2011* (2011).

[64] Gerland, Patrick, Raftery, Adrian E, Sevčíková, Hana, Li, Nan, Gu, Danan, Spoorenberg, Thomas, Alkema, Leontine, Fosdick, Bailey K, Chunn, Jennifer, Lalic, Nevena, Bay, Guiomar, Buettner, Thomas, Heilig, Gerhard K, and Wilmoth, John. World population stabilization unlikely this century. *Science 346*, 6206 (Oct. 2014), 234–7.

[65] Ginsberg, Jeremy, Mohebbi, Matthew H., Patel, Rajan S., Brammer, Lynnette, Smolinski, Mark S., and Brilliant, Larry. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (Feb. 2009), 1012–1014.

[66] Gneiting, Tilmann, and Raftery, Adrian E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 477 (Mar. 2007), 359–378.

[67] Goldberger, Arthur S. Structural equation methods in the social sciences. *Econometrica 40* (1972), 979–1001.

[68] Goodwin, Paul, Goodwin, and Paul. The holt-winters approach to exponential smoothing: 50 years old and going strong. *Foresight: The International Journal of Applied Forecasting*, 19 (2010), 30–33.

[69] Graefe, Andreas. German election forecasting: Comparing and combining methods for 2013. *German Politics 24*, 2 (Apr. 2015), 195–204.

[70] Green, A. J. Mass/length residuals: Measures of body condition or generators of spurious results? *Ecology 82*, 5 (2001), 1473–1483.

[71] Grossman, Jason, and Mackenzie, Fiona J. The randomized controlled trial: gold standard, or merely standard? *Perspectives in Biology and Medicine 48*, 4 (2005), 516–534.

[72] Guan, Peng, Huang, De-Sheng, and Zhou, Bao-Sen. Forecasting model for the incidence of hepatitis a based on artificial neural network. *World journal of gastroenterology 10*, 24 (Dec. 2004), 3579–82.

[73] Gubler, Duane J, Vasilakis, Nikos, and Musso, Didier. History and emergence of zika virus. *The Journal of Infectious Diseases 216*, suppl_10 (Dec. 2017), S860–S867.

[74] Hadfield, James, Megill, Colin, Bell, Sidney M, Huddleston, John, Potter, Barney, Callender, Charlton, Sagulenko, Pavel, Bedford, Trevor, and Neher, Richard A. Nextstrain: real-time tracking of pathogen evolution. *bioRxiv* (Nov. 2017), 224048.

[75] Haemig, Paul D, Sjöstedt de Luna, S, Grafström, A, Lithner, Stefan, Lundkvist, Åke, Waldenström, Jonas, Kindberg, Jonas, Stedt, Johan, and Olsén, Björn. Forecasting risk of tick-borne encephalitis (tbe): using data from wildlife and climate to predict next year's number of human victims. *Scandinavian journal of infectious diseases 43*, 5 (2011), 366–72.

[76] Hahn, Jinyong. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*, 2 (Mar. 1998), 315–331.

[77] Hanley, James A., and McNeil, Barbara J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology 143*, 1 (1982), 29–36.

[78] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* No. 2. Springer, 2009.

[79] Hayes, Richard J., and Moulton, Lawrence H. *Cluster Randomised Trials*, second ed. CRC Press, Boca Raton, Florida, USA, 2017.

[80] Held, Leonhard, Höhle, Michael, and Hofmann, Mathias. A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling: An International Journal 5*, 3 (Oct. 2005), 187–199.

[81] Held, Leonhard, Meyer, Sebastian, and Bracher, Johannes. Probabilistic forecasting in infectious disease epidemiology: the 13th armitage lecture. *Statistics in Medicine 36*, 22 (Sept. 2017), 3443–3460.

[82] Hersbach, Hans. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*, 5 (Oct. 2000), 559–570.

[83] Hii, Yien Ling, Zhu, Huaiping, Ng, Nawi, Ng, Lee Ching, and Rocklöv, Joacim. Forecast of dengue incidence using temperature and rainfall. *PLOS Neglected Tropical Diseases 6*, 11 (Nov. 2012), e1908.

[84] Hill, Austin B. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine 58*, 5 (May 1965), 295–300.

[85] Höhle, Michael, and an der Heiden, Matthias. Bayesian nowcasting during the stec o104:h4 outbreak in germany, 2011. *Biometrics 70*, 4 (Dec. 2014), 993–1002.

[86] Holt, Charles C. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting 20*, 1 (Jan. 2004), 5–10.

[87] Huber, John H, Childs, Marissa L, Caldwell, Jamie M, and Mordecai, Erin A. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and zika transmission. *bioRxiv* (Dec. 2017), 230383.

[88] Hyndman, Rob J., and Koehler, Anne B. Another look at measures of forecast accuracy. *International Journal of Forecasting 22*, 4 (2006), 679–688.

[89] Iannone, Richard. *stationaRy: Get Hourly Meteorological Data from Global Stations*, 2015. R package version 0.4.1.

[90] Jajosky, Ruth Ann, and Groseclose, Samuel L. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. *BMC Public Health 4*, 1 (Dec. 2004), 29.

[91] Jakob, Elizabeth M., Marshall, Samuel D., and Uetz, George W. Estimating fitness: A comparison of body condition indices. *Oikos 77*, 1 (Oct. 1996), 61.

[92] Johansson, Michael A., Dominici, Francesca, and Glass, Gregory E. Local and global effects of climate on dengue transmission in puerto rico. *PLOS Neglected Tropical Diseases 3*, 2 (2009), e382.

[93] Johansson, Michael A, et al. Advancing probabilistic epidemic forecasting through an open challenge: The dengue forecasting project. Under review.

[94] Johansson, Michael A., Mier-y Teran-Romero, Luis, Reefhuis, Jennita, Gilboa, Suzanne M., and Hills, Susan L. Zika and the risk of microcephaly. *New England Journal of Medicine 375*, 1 (July 2016), 1–4.

[95] Johnson, Leah R., Gramacy, Robert B., Cohen, Jeremy, Mordecai, Erin, Murdock, Courtney, Rohr, Jason, Ryan, Sadie J., Stewart-Ibarra, Anna M., and Weikel, Daniel. Phenomenological forecasting of disease incidence using heteroskedastic gaussian processes: a dengue case study.

[96] Jolliffe, Ian T., and Stephenson, David B. Introduction. In *Forecast verification : a practitioner's guide in atmospheric science*, Ian T. Jolliffe and David B. Stephenson, Eds. John Wiley & Sons Ltd, Chichester, West Sussex, England, 2003, pp. 1–12.

[97] Juliano, Steven A., O'Meara, George F., Morrill, Jeneen R., and Cutwa, Michele M. Desiccation and thermal tolerance of eggs and the coexistence of competing mosquitoes. *Oecologia 130*, 3 (Feb. 2002), 458–469.

[98] Kalayanarooj, Siripen. Standardized clinical management: Evidence of reduction of dengue haemorrhagic fever case-fatality rate in thailand. *Dengue Bulletin 23* (1999), 10–17.

[99] Kandula, Sasikiran, Yang, Wan, and Shaman, Jeffrey. Type- and subtype-specific influenza forecast. *American Journal of Epidemiology 185*, 5 (Mar. 2017), 395–402.

[100] Kane, Michael J, Price, Natalie, Scotch, Matthew, and Rabinowitz, Peter. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics 15*, 1 (2014), 276.

[101] Keegan, Lindsay T, Lessler, Justin, and Johansson, Michael A. Quantifying zika: Advancing the epidemiology of zika with quantitative models. *The Journal of Infectious Diseases 216*, suppl_10 (Dec. 2017), S884–S890.

[102] Keeling, Matthew James, and Rohani, Pejman. *Modeling Infectious Diseases in Humans and Animals*, vol. 47. Princeton University Press, Princeton, New Jersey, 2007.

[103] Kennedy, Edward H. Semiparametric theory.

[104] Kennedy, Edward H. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association* (Jan. 2018), 0–0.

[105] Kermack, W. O., and McKendrick, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 115*, 772 (1927), 700–721.

[106] Kleber de Oliveira, Wanderson, Cortez-Escalante, Juan, De Oliveira, Wanessa Tenório Gonçalves Holanda, do Carmo, Greice Madeleine Ikeda, Henriques, Cláudio Maierovitch Pessanha, Coelho, Giovanini Evelim, and Araújo de França, Giovanny Vinícius. Increase in reported prevalence of microcephaly in infants born to women living in areas with confirmed zika virus transmission during the first trimester of pregnancy — brazil, 2015. *MMWR. Morbidity and Mortality Weekly Report 65*, 9 (Mar. 2016), 242–247.

[107] Kotiaho, Janne S., Simmons, Leigh W., and Tomkins, Joseph L. Towards a resolution of the lek paradox. *Nature 410*, 6829 (Apr. 2001), 684–686.

[108] Lauer, S.A., Sakrejda, K., Ray, E.L., Keegan, L.T., Bi, Q., Suangtho, P., Hinjoy, S., Iamsirithaworn, S., Suthachana, S., Laosiritaworn, Y., Cummings, D.A.T., Lessler, J., and Reich, N.G. Prospective forecasts of annual dengue hemorrhagic fever incidence in thailand, 2010–2014. *Proceedings of the National Academy of Sciences of the United States of America 115*, 10 (2018).

[109] Lazer, D., Kennedy, R., King, G., and Vespignani, A. The parable of google flu: Traps in big data analysis. *Science 343*, 6176 (Mar. 2014), 1203–1205.

[110] Lega, Joceline, and Brown, Heidi E. Data-driven outbreak forecasting with a simple nonlinear growth model. *Epidemics 17* (Dec. 2016), 19–26.

[111] Lessler, Justin, and Cummings, Derek A. T. Mechanistic models of infectious disease and their impact on public health. *American Journal of Epidemiology 183*, 5 (Mar. 2016), 415–422.

[112] Lewis-Beck, Michael S., and Stegmaier, Mary. Us presidential election forecasting. *PS: Political Science & Politics 47*, 02 (Apr. 2014), 284–288.

[113] Liu, Hai-Ning, Gao, Li-Dong, Chowell, Gerardo, Hu, Shi-Xiong, Lin, Xiao-Ling, Li, Xiu-Jun, Ma, Gui-Hua, Huang, Ru, Yang, Hui-Suo, Tian, Huaiyu, and Xiao, Hong. Time-specific ecologic niche models forecast the risk of hemorrhagic fever with renal syndrome in dongting lake district, china, 2005–2010. *PLoS ONE 9*, 9 (Sept. 2014), e106839.

[114] Lowe, Rachel, Bailey, Trevor C., Stephenson, David B., Graham, Richard J., Coelho, Caio A. S., Sá Carvalho, Marilia, and Barcellos, Christovam. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in brazil. *Computers and Geosciences 37*, 3 (2011), 371–381.

[115] Lowe, Rachel, Barcellos, Christovam, Coelho, Caio A S, Bailey, Trevor C, Coelho, Giovanini Evelim, Graham, Richard, Jupp, Tim, Ramalho, Walter Massa, Carvalho, Marilia Sá, Stephenson, David B, and Rodó, Xavier. Dengue outlook for the world cup in brazil: an early warning model framework driven by real-time seasonal climate forecasts. *The Lancet Infectious Diseases 14*, 7 (July 2014), 619–626.

[116] Lowe, Rachel, Cazelles, Bernard, Paul, Richard, and Rodó, Xavier. Quantifying the added value of climate information in a spatio-temporal dengue model. *Stochastic Environmental Research and Risk Assessment 30*, 8 (Dec. 2016), 2067–2078.

[117] Lowe, Rachel, Coelho, Caio AS, Barcellos, Christovam, Carvalho, Marilia Sá, Catão, Rafael De Castro, Coelho, Giovanini E., Ramalho, Walter Massa, Bailey, Trevor C., Stephenson, David B., and Rodó, Xavier. Evaluating probabilistic dengue risk forecasts from a prototype early warning system for brazil. *eLife 5*, FEBRUARY2016 (Feb. 2016), e11285.

[118] Lowe, Rachel, Stewart-Ibarra, Anna M, Petrova, Desislava, García-Díez, Markel, Borbor-Cordova, Mercy J, Mejía, Raúl, Regato, Mary, and Rodó, Xavier. Climate services for health: predicting the evolution of the 2016 dengue season in machala, ecuador. *The Lancet Planetary Health 1*, 4 (July 2017), e142–e151.

[119] Lu, Fred Sun, Hou, Suqin, Baltrusaitis, Kristin, Shah, Manan, Leskovec, Jure, Sosic, Rok, Hawkins, Jared, Brownstein, John, Conidi, Giuseppe, Gunn, Julia, Gray, Josh, Zink, Anna, and Santillana, Mauricio. Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the boston metropolis. *JMIR public health and surveillance 4*, 1 (Jan. 2018), e4.

[120] Lu, Hsin-Min, Zeng, Daniel, and Chen, Hsinchun. Prospective infectious disease outbreak detection using markov switching models. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 22*, 4 (2010), 565–577.

[121] Magidson, Jay. Correlated component regression: Re-thinking regression in the presence of near collinearity. In *New Perspectives in Partial Least Squares and Related Methods*, H Abdi, W Chin, V Esposito Vinzi, G Russolillo, and L Trinchera, Eds., vol. 56. Springer, New York, New York, USA, 2013, pp. 65–78.

[122] Marshall, Geoffrey, Blacklock, JWS, Cameron, C, Capon, NB, Cruickshank, R, Gaddum, JH, Heaf, FRG, Hill, AB, Houghton, LE, Hoyle, JC, and Raistrick, H. Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal 2* (1948), 769–782.

[123] McCloskey, Donald N. The art of forecasting, ancient to modern times. *Cato Journal 12*, 1 (1992), 23–48.

[124] McGough, Sarah F., Brownstein, John S., Hawkins, Jared B., and Santillana, Mauricio. Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data. *PLOS Neglected Tropical Diseases 11*, 1 (Jan. 2017), e0005295.

[125] McGowan, Craig J., Biggerstaff, Matthew, Johansson, Michael, Apfeldorf, Karyn M., Ben-Nun, Michal, Brooks, Logan, Convertino, Matteo, Erraguntla, Madhav, Farrow, David C., Freeze, John, Ghosh, Saurav, Hyun, Sangwon, Kandula, Sasikiran, Lega, Joceline, Liu, Yang, Michaud, Nicholas, Morita, Haruka, Niemi, Jarad, Ramakrishnan, Naren, Ray, Evan L., Reich, Nicholas G., Riley, Pete, Shaman, Jeffrey, Tibshirani, Ryan, Vespignani, Alessandro, Zhang, Qian, and Reed, Carrie. Collaborative efforts to forecast seasonal influenza in the united states, 2015–2016. *Scientific Reports 9*, 1 (Dec. 2019), 683.

[126] Menne, Matthew J., Durre, Imke, Korzeniewski, Bryant, McNeal, Shelley, Thomas, Kristy, Yin, Xungang, Anthony, Steven, Ray, Ron, Vose, Russell S., E.Gleason, Byron, and Houston, Tamara G. Global Historical Climatology Network-Daily (GHCN-Daily), Version 3, Thailand weather stations, 2012. Access date: May 5, 2015.

[127] Menne, Matthew J., Durre, Imke, Vose, Russell S., Gleason, Byron E., and Houston, Tamara G. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology 29*, 7 (July 2012), 897–910.

[128] Merrill, Ray. Historic developments in epidemiology. In *Introduction to Epidemiology*. Jones & Bartlett Learning, 2010, ch. 2, pp. 23–46.

[129] Meyer, Sebastian, Held, Leonhard, and Höhle, Michael. Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *Journal of Statistical Software 77*, 11 (2017).

[130] Milham, Willis Isbister. *Meteorology: a text-book on the weather, the causes of its changes, and weather forecasting for the student and general reader*. Norwood The Macmillan Company, New York, 1918.

[131] Moore, K. L., and van der Laan, M. J. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine 28*, 1 (Jan. 2009), 39–64.

[132] Moore, Sean M., Monaghan, Andrew, Griffith, Kevin S., Apangu, Titus, Mead, Paul S., and Eisen, Rebecca J. Improvement of disease prediction and modeling through the use of meteorological ensembles: Human plague in uganda. *PLoS ONE 7*, 9 (Sept. 2012), e44431.

[133] Moran, Kelly R., Fairchild, Geoffrey, Generous, Nicholas, Hickmann, Kyle, Osthus, Dave, Priedhorsky, Reid, Hyman, James, and Del Valle, Sara Y. Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast. *Journal of Infectious Diseases 214*, suppl 4 (Dec. 2016), S404–S408.

[134] Morris, Dylan H, Gostic, Katelyn M, Pompei, Simone, Bedford, Trevor, Łuksza, Marta, Neher, Richard A, Grenfell, Bryan T, Lässig, Michael, and Mccauley, John W. Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends in Microbiology* (2017).

[135] Murphy, Allan H. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting 8*, 2 (June 1993), 281–293.

[136] Musso, Didier, and Gubler, Duane J. Zika virus. *Clinical microbiology reviews 29*, 3 (July 2016), 487–524.

[137] Myers, M F, Rogers, D J, Cox, J, Flahault, A, and Hay, S I. Forecasting disease risk for increased epidemic preparedness in public health. *Advances in parasitology 47* (2000), 309–30.

[138] National Oceanographic and Atmospheric Administration. Dengue forecasting project website, June 2015.

[139] Newey, Whitney K. Semiparametric efficiency bounds. *Journal of Applied Econometrics 5*, 2 (Apr. 1990), 99–135.

[140] Neyman, Jerzy. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Statistical Science 5* (1923), 465–480.

[141] Ng, Andrew Y. Preventing "overfitting" of cross-validation data. In *In Proceedings of the Fourteenth International Conference on Machine Learning* (San Francisco, CA, USA, 1997), Douglas H. Fisher, Ed., Morgan Kaufmann, pp. 245–253.

[142] Nishiura, Hiroshi. Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (h1n1-2009). *BioMedical Engineering OnLine 10*, 1 (2011), 15.

[143] Nurnchut, Auttawit, Aye, Yin Myo, Sookvech, Supaporn, Sookkhum, Sanya, Suwannachairob, Aorathai, Buathong, Rome, PipatJaturon, Natthakij, Cokthong, Vathin, Prommong, Nattasis, Saitaya, Supapich, Chenyawanich, Kitti, Suwanpatoomlert, Sarawoot, Jarasarn, Yoowarat, Chairangab, Todsaporn, Jeerasith, Wannakorn, Kithtiwiroch, Ithi, and Kithtiwiroch, Nikhom. A cluster of suspected cases of zika leading to uncommon dengue serotypes with possible coexisting zika virus in northern thailand, 2016. *OSIR Journal 11*, 3 (Sept. 2018), 13–21.

[144] Olson, Donald R., Konty, Kevin J., Paladini, Marc, Viboud, Cecile, and Simonsen, Lone. Reassessing google flu trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology 9*, 10 (Oct. 2013), e1003256.

[145] Osthus, Dave, Hickmann, Kyle S, Caragea, Petruţa C, Higdon, Dave, and Del Valle, Sara Y. Forecasting seasonal influenza with a state-space sir model. *The annals of applied statistics 11*, 1 (Mar. 2017), 202–224.

[146] Papanikolaou, Panagiotis N, Christidi, Georgia D, and Ioannidis, John P A. Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne 174*, 5 (Feb. 2006), 635–41.

[147] Pearl, Judea. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1988.

[148] Pearl, Judea. *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, New York, NY, New York, 2009.

[149] Pearl, Judea. An introduction to causal inference. *The international journal of biostatistics 6*, 2 (Feb. 2010), Article 7.

[150] Petersen, Maya L, Porter, Kristin E, Gruber, Susan, Wang, Yue, and Van Der Laan, Mark J. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research 21*, 1 (2010), 31–54.

[151] Petersen, Maya L, and van der Laan, Mark J. Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.) 25*, 3 (May 2014), 418–26.

[152] Priedhorsky, Reid, Osthus, Dave, Daughton, Ashlynn R., Moran, Kelly R., and Culotta, Aron. Deceptiveness of internet data for disease surveillance.

[153] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[154] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[155] Raftery, Adrian E, Li, Nan, Ševčíková, Hana, Gerland, Patrick, and Heilig, Gerhard K. Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences of the United States of America 109*, 35 (Aug. 2012), 13915–21.

[156] Ray, Evan L., and Reich, Nicholas G. Prediction of infectious disease epidemics via weighted density ensembles. *arXiv* (Mar. 2017).

[157] Reich, Nicholas G, Brooks, Logan C, Fox, Spencer J, Kandula, Sasikiran, McGowan, Craig J, Moore, Evan, Osthus, Dave, Ray, Evan L, Tushar, Abhinav, Yamana, Teresa K, Biggerstaff, Matthew, Johansson, Michael A, Rosenfeld, Roni, and Shaman, Jeffrey. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences of the United States of America* (Jan. 2019).

[158] Reich, Nicholas G., Lauer, Stephen A., Sakrejda, Krzysztof, Iamsirithaworn, Sopon, Hinjoy, Soawapak, Suangtho, Paphanij, Suthachana, Suthanun, Clapham, Hannah E., Salje, Henrik, Cummings, Derek A. T., and Lessler, Justin. Challenges in real-time prediction of infectious disease: A case study of dengue in thailand. *PLOS Neglected Tropical Diseases 10*, 6 (June 2016), e0004761.

[159] Reich, Nicholas G., Lessler, Justin, Sakrejda, Krzysztof, Lauer, Stephen A., Iamsirithaworn, Sopon, and Cummings, Derek A. T. Case study in evaluating time series prediction models using the relative mean absolute error. *The American Statistician 70*, 3 (Feb. 2016), 285–292.

[160] Reich, Nicholas G., Shrestha, Sourya, King, Aaron A., Rohani, Pejman, Lessler, Justin, Kalayanarooj, Siripen, Yoon, In-Kyu, Gibbons, Robert V., Burke, Donald S., and Cummings, Derek A. T. Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *Journal of the Royal Society, Interface 10*, 86 (2013), 20130414.

[161] Rigau-Pérez, José G., Clark, Gary G., Gubler, Duane J., Reiter, Paul, Sanders, Eduard J., and Vorndam, A. Vance. Dengue and dengue haemorrhagic fever. *Lancet 352*, 9132 (1998), 971–977.

[162] Robins, James. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling 7*, 9-12 (1986), 1393–1512.

[163] Robins, James. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases 40*, Supplement 2 (1987), 139S–161S.

[164] Robins, James M, Hernán, Miguel Ángel, and Brumback, Babette. Marginal structural models and causal inference in epidemiology. *Epidemiology 11*, 5 (2000), 550–560.

[165] Robins, James M., Rotnitzky, Andrea, and Zhao, Lue Ping. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*, 427 (Sept. 1994), 846–866.

[166] Rosenbaum, Paul R. Observational studies. Springer, New York, NY, 2002, pp. 1–17.

[167] Rosenbaum, Paul R., and Rubin, Donald B. The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 1 (Apr. 1983), 41–55.

[168] Rosenblum, Michael, and van der Laan, Mark J. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics 6*, 1 (Apr. 2010), Article 13.

[169] Rutvisuttinunt, Wiriya, Fernandez, Stefan, Thaisomboonsuk, Butsaya, Yoon, In-Kyu, Chinnawirotpisan, Piyawan, Klungthong, Chonticha, Plipat, Tanarak, Nisalak, Ananda, Hermann, Laura, Manasatienkij, Wudtichai, Buathong, Rome, and Akrasewi, Passakorn. Detection of zika virus infection in thailand, 2012–2014. *The American Journal of Tropical Medicine and Hygiene 93*, 2 (Aug. 2015), 380–383.

[170] Salathé, Marcel. Digital epidemiology: what is it, and where is it going? *Life Sciences, Society and Policy 14*, 1 (Dec. 2018), 1.

[171] Sánchez-Chardi, Alejandro, Peñarroja-Matutano, Cristina, Ribeiro, Ciro Alberto Oliveira, and Nadal, Jacint. Bioaccumulation of metals and effects of a landfill in small mammals. part ii. the wood mouse, apodemus sylvaticus. *Chemosphere 70*, 1 (Nov. 2007), 101–109.

[172] Santillana, Mauricio, Zhang, D Wendong, Althouse, Benjamin M, and Ayers, John W. What can digital disease detection learn from (an external revision to) google flu trends? *American journal of preventive medicine 47*, 3 (Sept. 2014), 341–7.

[173] Scharfstein, Daniel O, Rotnitzky, Andrea, and Robins, James M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94* (1999), 1096–1120.

[174] Scott, Thomas W., Morrison, Amy C., Lorenz, Leslie H., Clark, Gary G., Strickman, Daniel, Kittayapong, Pattamaporn, Zhou, Hong, and Edman, John D. Longitudinal studies of aedes aegypti (diptera: Culicidae) in thailand and puerto rico : Population dynamics. *Journal Medical Entomology 37*, 1 (Jan. 2000), 77–88.

[175] Shaman, Jeffrey, and Karspeck, Alicia. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences of the United States of America 109*, 3 (2012), 20425–30.

[176] Shaman, Jeffrey, Karspeck, Alicia, Yang, Wan, Tamerius, James, and Lipsitch, Marc. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications 4* (Dec. 2013), 2837.

[177] Shaman, Jeffrey, and Kohn, Melvin. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences of the United States of America 106*, 9 (Mar. 2009), 3243–8.

[178] Shaman, Jeffrey, Yang, Wan, and Kandula, Sasikiran. Inference and forecast of the current west african ebola outbreak in guinea, sierra leone and liberia. *PLoS Currents 6* (2014), ecurrents.outbreaks.3408774290b1a0f2dd7cae877c8b8f.

[179] Shao, Jun. Linear model selection by cross-validation. *Journal of the American Statistical Association 88*, 422 (1993), 486–494.

[180] Shen, Changyu, Li, Xiaochun, and Li, Lingling. Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in medicine 33*, 4 (Feb. 2014), 555–68.

[181] Siettos, Constantinos I, and Russo, Lucia. Mathematical modeling of infectious disease dynamics. *Virulence 4*, 4 (May 2013), 295–306.

[182] Silver, Nate. *The signal and the noise: why so many predictions fail but some don't*. The Penguin Group, New York, New York, USA, 2012.

[183] Snow, John. *On the mode of communication of cholera*, 2nd ed. Reprinted by Commonwealth Fund, 1936, New York, 1855.

[184] Soyiri, Ireneous N., and Reidpath, Daniel D. An overview of health forecasting. *Environmental Health and Preventive Medicine 18*, 1 (Jan. 2013), 1–9.

[185] Stanaway, Jeffrey D., Shepard, Donald S., Undurraga, Eduardo A., Halasa, Yara A., Coffeng, Luc E., Brady, Oliver J., Hay, Simon I., Bedi, Neeraj, Bensenor, Isabela M., Castañeda-Orjuela, Carlos A., Chuang, Ting Wu, Gibney, Katherine B., Memish, Ziad A., Rafay, Anwar, Ukwaja, Kingsley N., Yonemoto, Naohiro, and Murray, Christopher J. L. The global burden of dengue: an analysis from the global burden of disease study 2013. *The Lancet Infectious Diseases 16*, 6 (2016), 712–723.

[186] Stettler, Karin, Beltramello, Martina, Espinosa, Diego A, Graham, Victoria, Cassotta, Antonino, Bianchi, Siro, Vanzetta, Fabrizia, Minola, Andrea, Jaconi, Stefano, Mele, Federico, Foglierini, Mathilde, Pedotti, Mattia, Simonelli, Luca, Dowall, Stuart, Atkinson, Barry, Percivalle, Elena, Simmons, Cameron P, Varani, Luca, Blum, Johannes, Baldanti, Fausto, Cameroni, Elisabetta, Hewson, Roger, Harris, Eva, Lanzavecchia, Antonio, Sallusto, Federica, and Corti, Davide. Specificity, cross-reactivity, and function of antibodies elicited by zika virus infection. *Science (New York, N.Y.) 353*, 6301 (Aug. 2016), 823–6.

[187] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B 36*, 2 (1974), 111–147.

[188] Stroup, Donna F., Berlin, Jesse A., Morton, Sally C., Olkin, Ingram, Williamson, G. David, Rennie, Drummond, Moher, David, Becker, Betsy J., Sipe, Theresa Ann, Thacker, Stephen B., (MOOSE), and Group, for the Meta-analysis Of Observational Studies in Epidemiology. Meta-analysis of observational studies in epidemiology: A proposal for reporting. *JAMA 283*, 15 (Apr. 2000), 2008.

[189] Sumi, Ayako, and Kamo, Ken-ichi. Mem spectral analysis for predicting influenza epidemics in japan. *Environmental Health and Preventive Medicine 17*, 2 (Mar. 2012), 98–108.

[190] Surowiecki, James. *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations.* Doubleday, 2004.

[191] Teixeira, Maria da Glória, Barreto, Maurício L., Costa, Maria da Conceição N., Ferreira, Leila Denize A., Vasconcelos, Pedro F. C., and Cairncross, Sandy. Dynamics of dengue virus circulation: a silent epidemic in a complex urban area. *Tropical Medicine & International Health 7*, 9 (Sept. 2002), 757–762.

[192] Tsiatis, Anastasios A., Davidian, Marie, Zhang, Min, and Lu, Xiaomin. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine 27*, 23 (Oct. 2008), 4658–4677.

[193] Unkel, Steffen, Farrington, C. Paddy, Garthwaite, Paul H., Robertson, Chris, and Andrews, Nick. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 175*, 1 (Jan. 2012), 49–82.

[194] van der Laan, M J, and Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer, 2011.

[195] van der Laan, Mark J, and Gruber, Susan. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics 6*, 1 (May 2010), Article 17.

[196] van der Laan, Mark J., and Robins, James M. *Unified methods for censored longitudinal data and causality.* Springer, 2003.

[197] Venkatramanan, Srinivasan, Lewis, Bryan, Chen, Jiangzhuo, Higdon, Dave, Vullikanti, Anil, and Marathe, Madhav. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics 22* (Mar. 2018), 43–49.

[198] Viboud, Cécile, Sun, Kaiyuan, Gaffey, Robert, Ajelli, Marco, Fumanelli, Laura, Merler, Stefano, Zhang, Qian, and Chowell, Gerardo. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics 22* (Mar. 2018), 13–21.

[199] Viboud, Cécile, Sun, Kaiyuan, Gaffey, Robert, Ajelli, Marco, Fumanelli, Laura, Merler, Stefano, Zhang, Qian, Chowell, Gerardo, Simonsen, Lone, and Vespignani, Alessandro. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* (Aug. 2017).

[200] Waggoner, Jesse J., Balmaseda, Angel, Gresh, Lionel, Sahoo, Malaya K., Montoya, Magelda, Wang, Chunling, Abeynayake, Janaki, Kuan, Guillermina, Pinsky, Benjamin A., and Harris, Eva. Homotypic dengue virus reinfections in nicaraguan children. *Journal of Infectious Diseases 214*, 7 (Oct. 2016), 986–993.

[201] Wainwright, Martin J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*, 5 (May 2009), 2183–2202.

[202] Wearing, Helen J., and Rohani, Pejman. Ecological and immunological determinants of dengue epidemics. *Proceedings of the National Academy of Sciences 103*, 31 (Aug. 2006), 11802–11807.

[203] Wikan, Nitwara, Suputtamongkol, Yupin, Yoksan, Sutee, Smith, Duncan R., and Auewarakul, Prasert. Immunological evidence of zika virus transmission in thailand. *Asian Pacific Journal of Tropical Medicine 9*, 2 (Feb. 2016), 141–144.

[204] Winters, Peter R. Forecasting sales by exponentially weighted moving averages. *Management Science 6*, 3 (Apr. 1960), 324–342.

[205] Wongsawat, Jurai. Strategies for zika control in thailand, 2017.

[206] World Health Organization. Fact sheets: Infectious diseases.

[207] Wu, Pei-Chih, Guo, How-Ran, Lung, Shih-Chun, Lin, Chuan-Yao, and Su, Huey-Jen. Weather as an effective predictor for occurrence of dengue fever in taiwan. *Acta Tropica 103*, 1 (2007), 50–57.

[208] Xu, Qinneng, Gel, Yulia R., Ramirez Ramirez, L. Leticia, Nezafati, Kusha, Zhang, Qingpeng, and Tsui, Kwok-Leung. Forecasting influenza in hong kong with google search queries and statistical model fusion. *PLOS ONE 12*, 5 (May 2017), e0176690.

[209] Yamana, Teresa K., Kandula, Sasikiran, and Shaman, Jeffrey. Superensemble forecasts of dengue outbreaks. *Journal of The Royal Society Interface 13*, 123 (2016), 20160410.

[210] Yamana, Teresa K., Kandula, Sasikiran, and Shaman, Jeffrey. Individual versus superensemble forecasts of seasonal influenza outbreaks in the united states. *PLOS Computational Biology 13*, 11 (Nov. 2017), e1005801.

[211] Yan, Weirong, Xu, Yong, Yang, Xiaobing, and Zhou, Yikai. A hybrid model for short-term bacillary dysentery prediction in yichang city, china. *Japanese Journal of Infectious Diseases 63*, 4 (2010), 264–270.

[212] Yang, Shihao, Kou, Samuel C., Lu, Fred, Brownstein, John S., Brooke, Nicholas, and Santillana, Mauricio. Advances in using internet searches to track dengue. *PLOS Computational Biology 13*, 7 (July 2017), e1005607.

[213] Yang, Shihao, Santillana, Mauricio, and Kou, S C. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences of the United States of America 112*, 47 (Nov. 2015), 14473–8.

[214] Yoshioka, A. Use of randomisation in the medical research council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. *BMJ (Clinical research ed.) 317*, 7167 (Oct. 1998), 1220–3.